SPRINGER BRIEFS IN STATISTICS

Basilio de Bragança Pereira
Carlos Alberto de Bragança Pereira

Model Choice in Nonnested Families

Springer

SpringerBriefs in Statistics

More information about this series at http://www.springer.com/series/8921

Basilio de Bragança Pereira
Carlos Alberto de Bragança Pereira

Model Choice in Nonnested Families

Springer

Basilio de Bragança Pereira
COPPE
Federal University of Rio de Janeiro
Rio de Janeiro
Brazil

Carlos Alberto de Bragança Pereira
Department of Statistics
University of São Paulo
São Paulo
Brazil

ISSN 2191-544X ISSN 2191-5458 (electronic)
SpringerBriefs in Statistics
ISBN 978-3-662-53735-0 ISBN 978-3-662-53736-7 (eBook)
DOI 10.1007/978-3-662-53736-7

Library of Congress Control Number: 2016961264

Printed on acid-free paper

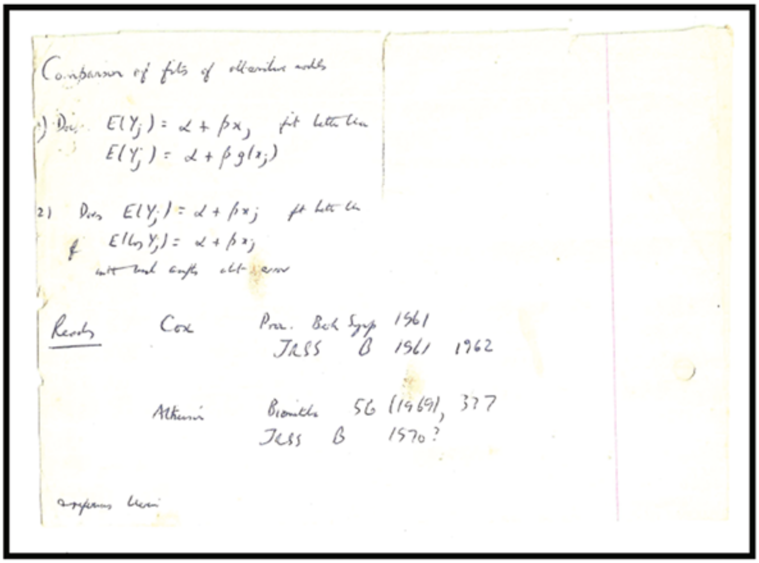
This Springer imprint is published by Springer Nature
The registered company is Springer-Verlag GmbH Germany
The registered company address is: Heidelberger Platz 3, 14197 Berlin, Germany

To Sir David Cox

Preface

Model choice is a subject that involves artistic and individual components that depend on the area of application, the amount of knowledge on the available models, and one's sense of aesthetics.

The first author's interest in model choice began in the summer of 1973: after a year of attending lectures at Imperial College, he made an appointment with his supervisor to decide his thesis topic. They summarized the meeting in the notes shown below. Although he did not pursue exactly the applications discussed at that time, he developed other results and applications concerning separate or nonnested model choice.



The beginning

Being a Bayesian, the second author began to be interested in the choice of models when solving a problem related to pollution in an industrial city in Brazil. He applied Bayesian significance tests to the mixture of models proposed by Cox instead of hypothesis tests and discrimination using Bayes factors.

Both authors have been following the advances in the subject, and this book is the result of their attempts to do so.

The authors are grateful to the writers and researchers on the subject from whom they have benefited and whom they have followed while writing this work, especially Mohammed Hashem Pesaran, and to Annibal P. SantAnna and Marlos Augusto G. Viana, who offered many suggestions for and corrections of the manuscript. The authors are also thankful for the many important contributions of Maria Ivanilde S. Araujo, Edilson F. de Arruda, Cachimo C. Assane, Rodrigo A. Collazo, Marcelo Lauretto, Brian A.R. de Melo, Fernando Poliano, and Julio Stern. They also thank Marcelo Fragoso and Augusto C.G. Vieira for the opportunity to complete their writing at Laboratório Nacional de Computação Científica–LNCC in Petrópolis, Brazil. Evelyn Best and Veronika Rosteck of Springer have been supportive and patient editors.

Petrópolis, Brazil
February 2016

Basilio de Bragança Pereira
Carlos Alberto de Bragança Pereira

Contents

Chapter 1
Preliminaries

Abstract In this chapter, the model choice problem is stated, and applications in several areas are presented. The definition of separate or nonnested models is given. The alternative approaches proposed by Cox (1961, 1962) for choosing among such models are presented. General references to the subject are mentioned, as are areas not covered in this book, namely, experimental design and discrepancy measures or information measures.

Keywords Bayes factors · Discrepancy measures · Discrimination · Hypothesis test · Likelihood ratio · Nonnested models · Separate models

1.1 Model Choice

In any scientific discipline, researchers constantly face the fundamental problem of choosing among alternative statistical models. In this context, the following questions arise (Atkinson 1970a; Claeskens and Hjort 2008):

(i) Is there evidence that the models produce significantly different fits to the data?
(ii) Assuming that one model is true, what is the evidence provided by the data that this model is really the true one?
(iii) If one model represents the currently maintained hypothesis, is there evidence of a departure from it in the direction of another model? If there is no maintained hypothesis, each model is on equal footing with every other model.
(iv) Models are approximations; therefore, it is more valuable to work with simpler models that are almost as good. We should keep in mind G.E.P. Box's maxim, "All models are wrong, but some are useful", and the "principle of parsimony" as expressed in the model formulation of Ockham's razor, "entities should not be multiplied without necessity". Approximate models share certain features with maps or dolls, for example. Maps fail to capture every detail of the landscape, just as dolls for children fail to capture every detail of the beings they represent, but both are useful. A surrealist view of this characteristic of models can be seen in the Magritte painting "The Treachery of Images" (1928–1929), in which he

B. de B. Pereira and C.A. de B. Pereira,
Model Choice in Nonnested Families, SpringerBriefs in Statistics,
DOI 10.1007/978-3-662-53736-7_1

painted a pipe and painted the following below it: “Ceci n’est pas une pipe”. When asked about the image, he replied, “Just try to fill it with tobacco”.

(v) All modeling is rooted in an appropriate context and its related objectives. Different schools of science may have different preferences. Breiman (2001) discusses the two cultures of statistics: the data-modeling culture (statistics: theory in search of data, or hypothesis-driven experiments Cox 2000) and the algorithm-modeling culture (data mining: data in search of a question or theory, or data-driven hypotheses Cox 2000). Thus, S. Karlin’s statement that “The purpose of models is not to fit the data, but to sharpen the question” (Claeskens and Hjort 2008, p. 2) contrasts with the black box view frequently adopted by the second culture, which is that a model is acceptable as long it works for prediction and classification. Different models may have different underlying physical or biological interpretations, even if they fit the data more or less equally well.

For the comparison of different models, the Neyman-Pearson theory of hypothesis testing or the Fisher theory of significance testing may be used if the models belong to the same family of distributions and if the relevant comparisons involve hierarchical (or nested) models.

However, special procedures are required if the models belong to families that are separate or nonnested in the sense that an arbitrary member of one family cannot be obtained as a limit of a model outside that family.

1.2 Types of Problems

Throughout this manuscript, Greek letters are used to denote unknown parameters. Suppose that the models under consideration are specified by the hypotheses H_f and H_g for densities $f(y, \alpha)$ and $g(y, \beta)$, respectively. The problems to be investigated in this book are illustrated using the following examples.

Example 1.1 Let $Y_1, \ldots, Y_n$ be independent and identically distributed (iid) random variables. Let H_f denote the hypothesis that their distribution function is lognormal with unknown parameter values, and let H_g denote the hypothesis that their distribution function is Weibull. Dumonceaux et al. (1973), Dumonceaux and Antle (1973) and Dumonceaux et al. (1973) have studied this problem.

Example 1.2 Let $Y_1, \ldots, Y_n$ be independent distributed random variables such that

$$\log Y_i = \mu + \sum_{r=1}^{m} z_{ir}\Theta_r + \log u_i,$$

where the z_i are m fixed regressors, μ is the general mean, and H_f and H_g specify alternative distributions for u_i, as in Example 1.1. Pereira (1978, 1981b) has studied this problem.

Example 1.3 The Pickering/Plat debate on the nature of hypertension is a widely published medical dispute. Plat claims that hypertension is a "disease" with underlying genetic determinants: one simply either has it or does not. He emphasizes that the skewness of the distribution of blood pressure is due to the effect of a dominant gene; thus, Plat espouses the hypothesis, denoted by H_f, that the blood pressure distribution is a mixture of two normal distributions.

Pickering argues that the designation "hypertension" is arbitrary and that the determinants of blood pressure are numerous and have small individual effects.

For Pickering, hypertension is not a disease but merely a label assigned to those with pressure readings in the upper tail of the distribution; thus, Pickering espouses the hypothesis, denoted by H_g, that the blood pressure distribution is a lognormal distribution. Refer to Shork et al. (1990) for details.

Example 1.4 Consider two alternative sets of covariates x and z for a regression problem and the alternative models

$$H_f : y_i = \alpha_0 + \sum_{r=1}^{\ell_1} x_{ij}\alpha_r + u_{if},$$

$$H_g : y_i = \beta_0 + \sum_{r=1}^{\ell_2} z_{ir}\beta_r + u_{ig},$$

where u_{if} and u_{ig} are (iid) random variables.

The problem of testing H_f against H_g has been addressed by Pesaran (1974) and Pereira (1984) under the assumptions that u_i follows a normal distribution and a Weibull distribution, respectively. Refer to Pereira (1981b, 1984) for an interesting result that emerges when there are alternative covariates and alternative distributions (Example 1.2). Practical applications include the following:

(i) discrimination between Constant Elasticity of Substitution (CES) and Variable Elasticity of Substitution (VES) production functions (Harvey 1977),
(ii) selection of level-differenced versus log-differenced stationary models (Pesaran and Pesaran 1995),
(iii) discrimination between monetarist and structuralist economic models for the Brazilian economy (Araujo and Pereira 2007), and
(iv) other empirical economic applications, as presented by McAleer (1995).

Example 1.5 The following alternative growth models have been considered for predicting AIDS cases in Brazil:

$$H_1 : \log y_t = \alpha_0 + \alpha_1 t + \alpha_3 Y_t,$$
$$H_2 : \log y_t = \beta_0 + \beta_2 \Delta \log Y_t + \beta_3 \log y_{t-1},$$
$$H_3 : \log y_t = \delta_0 + (\delta_1 Y_t + \delta_2 Y_t^2 + \delta_3 Y_3 \log Y_t),$$

where $y_t = \Delta Y_t = Y_t - Y_{t-1}$ and Y_t denotes cases of AIDS. These models were derived from those of Ord and Young (1988). The preferred model was found to be H_1, which includes the logistic, Gompertz and modified exponential growth models. The logistic model was ultimately chosen based on the confidence interval estimates (see Pereira and Migon 1989).

Example 1.6 Consider a time series Y_t. If the hypothesis of white noise properties is rejected, it might be interesting to test the following hypotheses:

$$H_f : y_t = \beta y_{t-1} + u_t \quad \text{against} \quad H_g : y_t = \varepsilon_t - \theta \varepsilon_{t-1},$$

where u_t and ε_t are iid normal random variables with means zero and variances $\tilde{\tau}_u^2$ and $\tilde{\tau}_\varepsilon^2$, respectively. These hypotheses are partially nonnested (Walker 1967).

Example 1.7 Consider binary observations Y with a covariate X and the hypotheses of a logistic or a probit model for these data, i.e.,

$$H_f : P(Y_i = 1) = \Phi(\alpha x_i) = \int_{-\infty}^{\alpha x_i} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\} dz,$$
$$H_g : P(y_i = 1) = \Lambda(\beta x_i) = \frac{e^{\beta x_i}}{1 + e^{\beta x_i}}.$$

Refer to Chambers and Cox (1967), Clarke and Signorino (1974), Morimune (1979), Pesaran and Pesaran (1993), Silva (2001), Genius and Strazzera (2002), and Monfardini (2003).

1.3 General Formulation

In this section, several methods first suggested by Cox in his original, fundamental paper (Cox 1961; see also Cox (1962, 2013)) are presented. These methods form the basis of most later developments on nonnested model choice. In fact, they comprise the core of this book.

Let Y be a vector of observations, and let H_f and H_g denote the hypotheses that the probability density function (p.d.f) of Y is $f(y, \alpha)$ or $g(y, \beta)$, respectively, where α and β are vectors of unknown parameters such that $\alpha \in \Omega_\alpha$ and $\alpha \in \Omega_\beta$, where Ω_α and Ω_β are the parameter spaces. It is also assumed that the families are separate in the sense defined above.

The formal definition of separate or nonnested models relies on the concept of discrepancies, such as the Ghosh and Subramanyam (1975) metric,

$$\begin{aligned} d(f, g) &= E_\alpha\{|f - g|\} \\ &= \int |f(y, \alpha) - g(y, \beta)| f(y, \alpha) dy, \end{aligned} \tag{1.1}$$

or the Kullback–Leibler divergence used by Pesaran (1987),

$$I(f, g) = E_\alpha\{\log f - \log g\} = E_\alpha\{\ell_{fg}\}$$
$$= \int_{\Omega_f} \log\{f(y, \alpha)/g(y, \beta)\} f(y, \alpha) dy. \tag{1.2}$$

Further possible metrics can be found in Linhart and Zucchini (1986).

Therefore, H_f and H_g are separate or nonnested if

$$\inf_{\Omega_f, \Omega_g} d(f, g) > 0 \text{ or}$$
$$\inf_{\Omega_f, \Omega_g} I(f, g) > 0.$$

Pesaran (1987) also defined partially nonnested hypotheses, for which the infimum is zero for some but not all of the parameters. Analogous expressions to (1.1) and (1.2) are defined when the roles of H_f and H_g are interchanged.

Several methods exist for addressing such model choice problems. Let us first consider a discrimination problem, where either H_f or H_g is true, and let us adopt the Bayesian approach.

Let π_f and π_g, such that $\pi_f + \pi_g = 1$, be the prior probabilities of H_f and H_g, respectively. $\pi_f(\alpha)$ and $\pi_g(\beta)$ are the prior probabilities for the parameters conditional on H_f and H_g, respectively. By Bayes' Theorem, the posterior odds ratio for H_f versus H_g is

$$\frac{\pi_f \int f(y, \alpha)\pi_f(\alpha)d\alpha}{\pi_g \int g(y, \beta)\pi_g(\beta)d\beta} = \frac{\pi_f}{\pi_g} B_{fg}(y). \tag{1.3}$$

The Bayes factor $B_{fg}(y)$ represents the weight of evidence provided by the data for H_f over H_g.

An alternative suggestion by Cox (1961) accounts for the losses $c_f(\alpha)$ and $c_g(\beta)$ incurred as a result of incorrectly rejecting H_f when α is the true parameter value or incorrectly rejecting H_g when β is the true parameter value, respectively. A decision theory approach leads to the following decision rule:

$$\pi_f \int_{\Omega_\alpha} f(y, \alpha)\pi_f(\alpha)c_f(\alpha)d\alpha \lesseqgtr \pi_g \int_{\Omega_\beta} f(y, \beta)\pi_g(\beta)c_g(\beta)d\beta. \tag{1.4}$$

Referring to Lindley (1961), Cox (1961) also developed the following large-sample approximation to (1.3) by expanding around the maximum likelihood values $\hat{\alpha}$ and $\hat{\beta}$:

$$\frac{f(y, \hat{\alpha})}{g(y, \hat{\beta})} \frac{\pi_f}{\pi_g} \frac{(2\pi)^{df/2}\pi_f(\hat{\alpha})}{(2\pi)^{dg/2}\pi_g(\hat{\beta})}, \frac{I_\alpha^{-1/2}}{I_\beta^{-1/2}}, \tag{1.5}$$

where df and dg are the numbers of dimensions of the parameters α and β and I_α and I_β are the information determinants for estimating α and β. For another approximation, refer to Cox and Hinkley (1978, p. 162).

If the prior distributions are available, then the Bayesian approach provides a general solution to the problem of discriminating between H_f and H_g. For the case in which the priors are unavailable, Cox (1961) suggested the introduction of the generalized Neyman–Pearson likelihood ratio

$$R_{fg} = e^{\hat{\ell}_{fg}} = \left\{ \frac{\sup\limits_{\Omega_\alpha} f(y, \alpha)}{\sup\limits_{\Omega_\beta} g(y, \beta)} \right\} = \frac{f(y, \hat{\alpha})}{g(y, \hat{\beta})} \tag{1.6}$$

as an alternative to (1.5), where R_{fg} is the log-likelihood ratio. A third suggestion was presented by Cox (1961) based on an examination of expression (1.6).

He noticed that an improper prior could not be used in (1.3), which is unspecified.

Cox (1961) went on to invoke the Obviously Arbitrary and Always Admissible (OAAAA method), suggested by Bernard (1959). It consists of three steps: taking a small number of points in Ω_α and Ω_β, evaluating the corresponding likelihood functions of these points under H_f and H_g, and computing the ratio of the average of the likelihood functions under H_f over the average of the likelihood functions under H_g. This corresponds to a Bayes solution with respect to the uniform prior over the two sets of points of the considered hypotheses. In fact, this method leads to a ratio of the mean likelihoods rather than a ratio of the maximum likelihoods, as in (1.6), and it is also related to the Bayesian procedures presented in Sects. 3.2.5 and 3.3 of Chap. 3.

For the case in which ℓ_{fg} is treated as a random variable denoted by L_{fg}, Cox (1961) presented several interpretations of the use of (1.6). Direct utilization of (1.6) is only meaningful if H_f and H_g specify simple hypotheses. In this case, it is sufficient to take the observed value of (1.6) to measure the evidence in favor of H_f. The same is not true if the numbers of parameters considered under H_f and H_g are different. In this case, one can always expect a better fit to the data using the model with more parameters when the other modeling aspects remain unchanged.

Considering the problem as one of significance testing, where the hypotheses H_f and H_g are considered in an asymmetrical rather than a discrimination manner, H_g represents the alternative for which a higher power is required. Cox (1961) suggestions for this case are based on the distribution of the statistic

$$T_f = \{\ell_f(\hat{\alpha}) - \ell_g(\hat{\beta})\} - E_{\hat{\alpha}}\{\ell_f(\hat{\alpha}) - \ell_g(\hat{\beta})\}, \tag{1.7}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the maximum likelihood estimators and $\ell_f(\alpha)$ and $\ell_g(\beta)$ are the log-likelihood functions under H_f and H_g, respectively. An analogous expression is obtained for T_g.

An alternative formulation presented by Cox (1961) considers an exponential mixture that includes the models corresponding to H_f and H_g as particular cases, where λ is a further unknown parameter:

$$\frac{f^{\lambda}(y,\alpha)g^{1-\lambda}(y,\beta)}{\int f^{\lambda}(y,\alpha)g^{1-\lambda}(y,\beta)dy}. \tag{1.8}$$

Here, the significances of $H_f : \lambda = 1$ and $H_f : \lambda = 0$ are tested.

Another comprehensive model is the linear mixture

$$\frac{\lambda f(y,\alpha) + (1-\lambda)g(y,\beta)}{\int[\lambda f(y,\alpha) + (1-\lambda)g(y,\beta)]dy}, \tag{1.9}$$

mentioned by Atkinson (1970b) and first studied by Quandt (1974).

Finally, a distinction should be drawn between discrimination and hypothesis testing. Discrimination begins with a given set of models, and the purpose is to select one of the models under consideration. By contrast, hypothesis testing asks whether there is statistically significant evidence of a departure from the null hypothesis in the direction of one or more alternative hypotheses. Rejection of the null hypothesis does not necessarily imply acceptance of any of the alternative hypotheses. In the case of separate hypothesis testing, it is possible that all models considered may be rejected or that all models may be accepted (not rejected).

1.4 Plan of the Book

Chapter 2 introduces the frequentist approach to the problem of testing separate models. A derivation of the Cox test is given. Alternative procedures are presented. The exponential mixture and its various econometric extensions are illustrated. False and nearest models and the related pseudo-maximum likelihood estimators are discussed. A comparison among alternative methods is briefly discussed in some cases.

Chapter 3 presents the Bayesian approach to the problem of discriminating among separate models. The limitations of Bayes factors are described, and alternative modified Bayes factors to resolve these limitations are presented. Bayesian significance testing is also presented.

Finally, Chap. 4 addresses the pure likelihood and support approaches as applied to certain data. Bootstrap and simulation approaches are also discussed.

Throughout the chapters, real-world examples and simulation results are presented and discussed to illustrate conceptual aspects.

Major areas that are not covered in this book include experimental design for the discrimination of alternative models and methods based on discrepancy and information measures, such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the Minimum Description Length (MDL), among others. Each of these topics is a subject of an entire book in itself.

1.5 Bibliographic Notes

The relevance of the topic of this book and its influence on later developments in statistics have recently been revisited by Cox (2013).

A brief history of the further work of Cox is provided in Araujo et al. (2005). Reviews and references of general interest can be found in Pereira (1977a), Pereira (1981d), Pereira (2005) and Pereira (2010). For regularity conditions for the Cox test, refer to White (1982) and also Pereira (1977b, 1981a).

In the 1980s, econometricians took great interest in this subject, which has been reviewed frequently: see MacKinnon (1983), McAleer and Pesaran (1986), McAleer (1987, 1995), Gourieroux and Monfort (1994), Szroeter (1999), Pesaran and Weeks (2001) and Pesaran and Ulloa (2008).

Bayesian statisticians in the 1990s developed alternative Bayes factors to overcome the difficulties related to the standard Bayes factor. Also of interest in the Bayesian context is the work of Poirer (1997) on the choice between two models when a third model is present in the background.

Finally, several references on areas not covered in this book are as follows: Alberton et al. (2011), for a recent study on experimental design, and Linhart and Zucchini (1986), Sakamoto et al. (1986), Burnham and Anderson (2002), Anderson (2008), Claeskens and Hjort (2008), Konishi and Kitagawa (2010), Rissanen (1989, 2010) and Wallace (2005), for discussions of discrepancy and information methods.

References

Alberton, A., Schwaab, M., Lobão, M.W.N., Pinto, J.C.: Experimental design for the joint model discrimination and precise parameter estimation through information measures. Chem. Eng. Sci. **66**, 1940–1952 (2011)

Anderson, D.R.: Model Based Inference in the Life Sciences: A Primer on Evidence. Springer, New York (2008)

Araujo, M.I., Pereira, B.B.: Bayes factors to discriminate separate multivariate regression using improper prior (in Portuguese). Revista Brasileira de Estatística **68**, 33–50 (2007)

Araujo, M.I., Pereira, B.B., Cleroux, R., Fernandes, M., Lazraq, A.: Separate families of models: sir David Cox contributions and recent developments. Student, **5**, 251–258 (2005)

Atkinson, A.C.: Posterior probabilities for choosing a regression model. Biometrika **65**, 39–48 (1970a)

Atkinson, A.C.: A method for discriminating between models (with discussion). J. R. Stat. Soc. B, 323–353 (1970b)

Bernard, G.A.: Control charts and stochastic processes (with discussion). J. R. Stat. Soc. **21**, 239–271 (1959)

Breiman, L.: Statistical modeling: the two cultures (with discussion). Stat. Sci. **16**, 199–231 (2001)

Burnham, K.P., Anderson, D.R.: Model Selection and Multimodel Inference: A Practical Information-Theoric Approach, 2nd edn. Springer, New York (2002)

Chambers, E.A., Cox, D.R.: Discrimination between alternative binary response data. Biometrika **54**, 573–578 (1967)

Clarke, K.A., Signorino, C.S.: Discriminating methods: tests for non-nested discrete choice models. Polit. Stud. **58**, 368–388 (1974)

Claeskens, G., Hjort, N.L.: Model Selection and Model Averaging. Cambridge University Press (2008)
Cox, D.R.: Tests of separate families of hypotheses. In: Proceedings 4th Berkeley Symposium in Mathematical Statistics and Probability, vol. 1, pp. 105–123 (1961). University of California Press
Cox, D.R.: Further results on test of separate families of hypotheses. J. R. Stat. Soc. B, 406–424 (1962)
Cox, D.R., Hinkley, D.V.: Problems and Solutions in Theoretical Statistics. Chapman and Hall (1978)
Cox, D.R.: Personal Communication, to B de B Pereira. Caxambu, MG (2000)
Cox, D.R.: A return to an old paper: tests of separate families of hypotheses (with discussion). J. R. Stat. Soc. B. **75**, 207–215 (2013)
Dumonceaux, R., Antle, C.E.: Discriminating between the logonormal and Weibull distribution. Technometrics **15**, 923–926 (1973)
Dumonceaux, R., Antle, C.E. adn Haas, G.: Likelihood ratio test for discriminating between two models with unknown location and scale parameters. Technometrics **15**, 19–927 (1973)
Genius, M., Strazzera, E.: A note about model selection and tests for non-nested contingent valuation models. Econ. Lett. **74**, 363–370 (2002)
Ghosh, J.K., Subramanyam, K.: Inference about separated families in large samples. Sankhyã: Ind. J. Stat. A **37**, 502–513 (1975)
Gourieroux, C., Monfort, A.: Testing non-nested hypotheses. In: Engle, R., McFadden, D.L. (eds) Handbook of North Holland, vol. IV, chapter 44, pp. 2585–2637 (1994)
Harvey, A.C.: Discrimination between CES and VES production function. Ann. Econ. Soc. Measur. **6**, 463–471 (1977)
Konishi, S., Kitagawa, G.: Information Criteria and Statistical Modeling. Springer, New York (2010)
Lindley, D.V.: The use of prior probability distributions in statistical inference. In: Proceedings 4th Berkeley Symposium in Mathematical Statistics and Probability, vol. 1, pp. 453–468 (1961). University of California Press
Linhart, H., Zucchini, W.: Model Selection. Wiley (1986)
MacKinnon, J.G.: Model specification tests against non-nested alternative (with discussion). Econ. Rev. **2**, 85–158 (1983)
McAleer, M.: Specification tests for separate models: a survey. In: King, M.L., Giles, D.E.A. (eds.) Specification Analysis in the Linear Model, pp. 146–195. Routledge and Kegan Paul (1987)
McAleer, M.: The significance of testing empirical non-nested models. J. Econ. **67**, 149–171 (1995)
McAleer, M., Pesaran, M.H.: Statistical inference in non-nested econometric models. Appl. Math. Comput. **20**, 271–311 (1986)
Monfardini, C.: An illustration of Cox's non-nested testing procedure for logit and probit models. Comput. Stat. Data Anal. **42**, 425–444 (2003)
Morimune, K.: Comparisons of normal and logistic models in the bivariate dichotomous analysis. Econometrica **47**, 957–975 (1979)
Ord, K., Young, P.: Time series models for technological forecasting (preliminar version). In: International Forecasting Symposium, 1989. Vancouver, Canada (1988)
Pereira, B.B.: Discriminating among separate models: A bibliography. Int. Stat. Rev. **45**, 163–172 (1977a)
Pereira, B.B.: A note on the consistency and on finite sample comparisons of some tests of separate families of hypotheses. Biometrika **64**, 109–113 (1977b)
Pereira, B.B.: Test of efficiencies of separate regression models. Biometrika **65**, 319–327 (1978)
Pereira, B.B.: Empirical comparisons of some tests of separate families of hypotheses. Metrika **25**, 219–234 (1981a)
Pereira, B.B.: Amendment: to tests and efficiencies of separate regression models. Biometrika **68**, 345 (1981b)
Pereira, B.B.: Choice of a survival model for patients with a brain tumour. Metrika **28**, 53–61 (1981c)

Pereira, B.B.: Discriminating among separate models: an additional bibliography. Int. Stat. Inf. **62**, 3, and In: Katti, S.K. (ed.) On the Preliminary Test for CEAS Model versus the Thompson Model for Predicting Soybean Production, Technical Report. Department of Statistics, University of Missouri, Columbia (1981d)

Pereira, B.B.: On the choice of a Weibull model. Estadistica J. Inter-Am. Stat. Inst. **26**, 157–163 (1984)

Pereira, B.B.: Separate families of hypotheses. In: Armitage, P., Calton, T. (eds.) Encyclopedia of Biostatistics, vol. 7, pp. 4881–4886, 2nd edn. Wiley (2005)

Pereira, B.B.: Tests for discriminating separate or non-nested models. In: Lovic, M. (ed.) International Encyclopedia of Statistical Science, vol. 3, pp. 1592–1595. Springer (2010)

Pereira, B.B., Migon, H.S.: Choice of models for predicting AIDS (in Portuguese). Atas da 3a. Escola de Séries Temporais e Econometria. ABE and FGV-Rio, pp. 13–20 (1989)

Pesaran, M.H.: On the general problem of model selection. Rev. Econ. Stud. **41**, 153–171 (1974)

Pesaran, M.H.: Pitfalls on testing non-nested hypotheses by the Lagrange multiplier method. J. Econ. **17**, 323–331 (1981)

Pesaran, M.H.: Global and partial nonnested hypothesis and asymptotic local power. Econ. Theory **3**, 69–97 (1987)

Pesaran, M.H., Pesaran, B.: A simulation approach to the problem of computing Cox's statistics for testing nonnested models. J. Econ. **57**, 377–392 (1993)

Pesaran, M.H., Ulloa: Non-nested hypothesis. In: Durlauf, S.N., Blume, L.E. (eds.) New Palgrave Dictionary of Economics, vol. 6, pp. 107–114. Palgrave MacMillan (2008)

Pesaran, B., Pesaran, M.H.: A non-nested test of level-differenced versus log-differenced stationary models. Econ. Rev. **14**, 213–227 (1995)

Pesaran, M H., Weeks, M.: Non-nested hypothesis testing: An overview. In: Baltagi, B.H. (ed.) Companion to Theoretical Econometrics. Basil Blackwell (2001)

Poirer, D.J.: Comparing and choosing between two models with a third model in the background. J. Econ. **78**, 139–151 (1997)

Quandt, R.E.: A comparison of methods for testing nonnested hypotheses. Rev. Econ. Stat. **56**, 92–99 (1974)

Rissanen, J.: Stochastic Complexity in Statistical Inquire. World Scientific Publishing (1989)

Rissanen, J.: Information and Complexity in Statistical Modeling. Springer (2010)

Sakamoto, Y., Ishiguro, M., Kitagawa, G.: Akaike Information Criterion Statistics. D Reidel Publishing Company (1986)

Shork, N.J., Weder, A.B., Shork, M.A.: On the asymmetry of biological frequency distributions. Genet. Epidemiol. **7**, 427–446 (1990)

Silva, J.M.C.: A score test for non-nested hypotheses with applications to discrete data models. J. Appl. Econ. **16**, 77–597 (2001)

Szroeter, J.: Testing non-nested econometric models. Curr. State Econ. Sci. **1**, 223–253 (1999)

Walker, A.M.: Some test of separate families of hypotheses in time series analysis. Biometrika **54**, 39–68 (1967)

Wallace, C.: Statistical and Inductive Inference by Minimum Message Length. Springer (2005)

White, A.M.: Regularity conditions for Cox's test of non-nested hypotheses. J. Econ. **19**, 301–318 (1982)

Zabel, J.E.: A comparison of nonnested tests for misspecified models using the method of approximate slopes. J. Econ. **57**, 205–232 (1993)

Chapter 2
Frequentist Methods

Abstract This chapter presents frequentist statistical methods. Hypothesis tests, namely, the Cox test and alternatives, are described. An interpretation of the test results is provided. Applications to the exponential, gamma, Weibull, and lognormal distributions are presented. Misspecification and the efficiencies of false regression models are studied. Certain properties of some of the procedures in terms of power and consistency are presented, both analytically and based on simulations. References to recent applications of the Cox test are mentioned, as is the relation of the pioneering work on the efficiency of false models to recent works on misspecification and what is known as the "Sandwich" formula for estimation of covariance.

Keywords Alternative hypothesis · Asymptotic power · Comprehensive models · Cox test · Exponential models · False models · Gamma models · Gradient test · Log-normal models · Neyman–Pearson likelihood ratio · Null hypothesis · Probability limit · Rao score test · Simulations · Wald test · Weibull models

2.1 Introduction

In Chap. 1 the key concepts related to choosing among separate models were discussed. The present chapter discusses frequentist solutions for solving this problem. Alternative tests are presented, along with some of their properties. The concepts of false models, pseudomaximum likelihood and misspecification are also discussed.

2.2 The Cox Test

2.2.1 Preliminaries

Let $y = (y_1, \ldots, y_n)$ be independent observations drawn from some unknown distribution F. Suppose that the null hypothesis $H_f : F \in \mathfrak{F}_f$ is to be tested, where $\mathfrak{F}_f$ is a family of probability distributions with density $f(y, \alpha)$ and α is an unknown vector parameter. Let a high power be required for the alternative hypothesis $H_g : F \in \mathfrak{F}_g$,

B. de B. Pereira and C.A. de B. Pereira,
Model Choice in Nonnested Families, SpringerBriefs in Statistics,
DOI 10.1007/978-3-662-53736-7_2

where $\mathfrak{F}_g$ is another family of probability distributions with density $g(y, \beta)$; here, β is an unknown vector parameter and $f(y, \alpha)$ and $g(y, \beta)$ are separate or nonnested models, as defined in Chap. 1.

The asymptotic test developed by Cox (1961, 1962) is based on a modification of the Neyman–Pearson maximum likelihood ratio. If H_f is the null hypothesis, then the considered test statistic is

$$T_{fg} = \ell_f(\hat{\alpha}) - \ell_g(\hat{\beta}) - E_{\hat{\alpha}}\left\{\ell_f(\hat{\alpha}) - \ell_g(\hat{\beta})\right\}, \tag{2.1}$$

as defined in Sect. 1.3. The following alternative interpretations and forms can also be used to compute this statistic, neither of which affects the null distribution under the null hypothesis (see Kent 1986 and his discussion of Cox 2013):

$$\begin{aligned} T_{fg} &= \ell_f(\hat{\alpha}) - \ell_g(\hat{\beta}) - E_{\hat{\alpha}}\left\{\ell_f(\alpha) - \ell_g(\beta_\alpha)\right\}, \\ T_{fg} &= \ell_f(\hat{\alpha}) - \ell_g(\hat{\beta}) - n \operatorname*{plim}_{n\to\infty}\left[n^{-1}\left\{\ell_f(\hat{\alpha}) - \ell_g(\hat{\beta})\right\}\right]_{\alpha=\hat{\alpha}}, \end{aligned} \tag{2.2}$$

where $\ell_f(\hat{\alpha})$ and $\ell_g(\hat{\beta})$ are the maximized log-likelihoods under H_f and H_g, respectively; $\hat{\alpha}$ and $\hat{\beta}$ denote the maximum likelihood estimators; β_α is the probability limit, as $n \longrightarrow \infty$, of $\hat{\beta}$ under H_f; plim represents convergence in probability; and the subscript α indicates that the means are calculated under H_f.

Because $\hat{\beta} \xrightarrow{p} \beta_\alpha$, we have

$$E_\alpha\left[\frac{\partial}{\partial\beta}\ell_g(\beta_\alpha)\right] = 0. \tag{2.3}$$

Example 2.1 (Cox 1961; Jackson 1968) The null hypothesis H_L is that the distribution is lognormal, and the alternative is that the distribution is exponential; that is,

$$\begin{aligned} &H_L : f_L(y, \alpha_1, \alpha_2) = y(2\pi\alpha_2)^{-1/2}\exp\left\{-(\log y - \alpha_1)^2/2\alpha_2\right\},\ \alpha = (\alpha_1, \alpha_2), \\ &H_E : f_E(y, \beta) = \exp(-y/\beta)/\beta. \end{aligned} \tag{2.4}$$

The maximum likelihood estimator is $\hat{\beta} = \bar{y}$. Under H_L,

$$\hat{\beta} \xrightarrow{p} \beta_\alpha = \exp\left\{\alpha_1 + \frac{1}{2}\alpha_2\right\} \text{ and}$$

$$\begin{aligned} \ell_g(\beta_\alpha) = \ell_E(\beta_{(\alpha_1,\alpha_2)}) &= \ln\left[\exp\left\{-(y/\mathrm{e}^{\alpha_1+\frac{1}{2}\alpha_2})\right\}/\mathrm{e}^{\alpha_1+\frac{1}{2}\alpha_2}\right] \\ &= -y/\mathrm{e}^{\alpha_1+\frac{1}{2}\alpha_2} - \alpha_1 - \frac{1}{2}\alpha_2 \end{aligned}$$

and

$$\frac{\partial}{\partial \alpha_1}\ell_E(\beta_{(\alpha_1,\alpha_2)}) = y/e^{\alpha_1+\frac{1}{2}\alpha_2} - 1,$$
$$\frac{\partial}{\partial \alpha_2}\ell_E(\beta_{(\alpha_1,\alpha_2)}) = y/e^{\alpha_1+\frac{1}{2}\alpha_2} - 1/2.$$

Therefore,

$$E_\alpha\left\{\frac{\partial}{\partial \alpha}\ell_E(\beta_{(\alpha_1,\alpha_2)})\right\} = (0, 0).$$

2.2.2 Remarks on the Distribution of T_{fg}

A heuristic general explanation of the distribution of the test statistic is presented below. A complete proof of the distributional properties and general regularity conditions for the Cox test are given in White (1982).

Expanding $\ell_f(\hat{\alpha})$, $\ell_g(\beta_\alpha)$, $E_{\hat{\alpha}}\left\{\ell_f(\alpha)\right\}$ and $E_{\hat{\alpha}}\left\{\ell_g(\beta_\alpha)\right\}$ around α and $\ell_g(\hat{\beta})$ around β, we obtain

$$\ell_f(\hat{\alpha}) \cong \ell_f(\alpha),\ \ell_f(\hat{\beta}) \cong \ell_g(\beta),$$
$$E_{\hat{\alpha}}\left\{\ell_f(\alpha)\right\} \cong E_\alpha\left\{\ell_f(\alpha)\right\} + (\hat{\alpha}-\alpha)' E_\alpha\left\{\ell_f(\alpha)\frac{\partial \ell_f(\alpha)}{\partial \alpha}\right\}$$

and

$$E_{\hat{\alpha}}\left\{\ell_g(\beta_\alpha)\right\} = E_\alpha\left\{\ell_g(\beta_\alpha)\right\} + (\hat{\alpha}-\alpha)' E_\alpha\left\{\ell_g(\beta_\alpha)\frac{\partial \ell_f(\alpha)}{\partial \alpha}\right\}.$$

Applying these results to Eq. (2.2), we obtain

$$\begin{aligned} T_{fg} &= \ell_f(\alpha) - \ell_g(\beta_\alpha) - E_\alpha\left\{\ell_f(\alpha) - \ell_g(\beta_\alpha)\right\} \\ &\quad -(\hat{\alpha}-\alpha)' E_\alpha\left\{\left(\ell_f(\alpha) - \ell_g(\beta_\alpha)\right)\frac{\partial \ell_f(\alpha)}{\partial \alpha}\right\}. \end{aligned} \tag{2.5}$$

By writing $Z = \ell_f(\alpha) - \ell_g(\beta) - E_\alpha\left\{\ell_f(\alpha) - \ell_g(\beta_\alpha)\right\}$ and using the fact that the asymptotic distribution of $\sqrt{n}(\hat{\alpha}-\alpha)$ is the same as that of $\sqrt{n}I^{-1}(\alpha)\frac{\partial \ell_f(\alpha)}{\partial \alpha}$, where $I(\alpha) = E_\alpha\left\{\frac{\partial \ell_f(\alpha)}{\partial \alpha}\right\}^2$, it follows that the variance is

$$\begin{aligned} V_\alpha(T_{fg}) &= V_\alpha\left(\ell_f(\alpha) - \ell_g(\beta_\alpha)\right) + E_\alpha\left\{Z\frac{\partial \ell_f(\alpha)}{\partial \alpha}\right\}' I^{-1}(\alpha) V\left(\frac{\partial \ell_f(\alpha)}{\partial \alpha}\right) I^{-1}(\alpha) \\ &\quad \times E_\alpha\left\{Z\frac{\partial \ell_f(\alpha)}{\partial \alpha}\right\} - 2E_\alpha\left\{Z\frac{\partial \ell_f(\alpha)}{\partial \alpha}\right\}' I(\alpha)^{-1} E\left\{Z\frac{\partial \ell_f(\alpha)}{\partial \alpha}\right\} \\ &= V_\alpha(Z) + Cov'\left\{Z\frac{\partial \ell_f(\alpha)}{\partial \alpha}\right\} I(\alpha)^{-1} Cov\left\{Z\frac{\partial \ell_f(\alpha)}{\partial \alpha}\right\}. \end{aligned} \tag{2.6}$$

Therefore, T_{fg} is the sum of the deviations of $\ell_f(\alpha) - \ell_g(\beta)$ from its regression on $\partial \ell_f(\alpha)/\partial(\alpha)$. Its order is $\sqrt{n}$ in probability, whereas the other terms are of order one in probability.

Expression (2.6) can be written as

$$V_\alpha(T_{fg}) = V_\alpha\left(\ell_f(\alpha) - \ell_g(\beta_\alpha)\right) - C'_\alpha I^{-1}(\alpha) C_\alpha, \tag{2.7}$$

where $C_\alpha = \frac{\partial}{\partial \alpha} E_\alpha \left\{\ell_f(\alpha) - \ell_g(\beta)\right\}$ and $I(\alpha)$ is the information matrix of α.

It also follows that (Cox 1961)

$$\begin{aligned}
Cov(\hat{\alpha}) &= -\tfrac{1}{n} E_\alpha \left(\tfrac{\partial^2 \ell_f(\alpha)}{\partial\alpha\partial\alpha}\right)^{-1}, \\
Cov(\hat{\alpha}, \hat{\beta}) &= \tfrac{1}{n} E_\alpha \left(\tfrac{\partial^2 \ell_f(\alpha)}{\partial\alpha\partial\alpha}\right)^{-1} \left(\tfrac{\partial\beta_\alpha}{\partial\alpha}\right), \\
Cov(\hat{\beta}) &= \tfrac{1}{n} \left\{ E_\alpha \left(\tfrac{\partial^2 \ell_g(\beta_\alpha)}{\partial\beta\partial\beta}\right)^{-1} E_\alpha \left(\tfrac{\partial\beta_\alpha}{\partial\beta}\right)' \left(\tfrac{\partial \ell_g(\beta_\alpha)}{\partial\beta}\right) E_\alpha \left(\tfrac{\partial^2 \ell_g(\beta_\alpha)}{\partial\beta\partial\beta}\right) \right\}.
\end{aligned} \tag{2.8}$$

T_{fg} is the sum of independent and identically distributed (iid) random variables with mean zero; therefore, quite generally, a strong central limit effect can be expected to apply, unless, of course, the individual components have a markedly badly behaved distribution.

2.2.3 *The Test Procedure*

When H_g is the null hypothesis and H_f is the alternative hypothesis, analogous results are obtained for a statistic T_{gf}. Because $C^*_{fg} = T_{fg}\left\{V(T_{fg})\right\}^{-1/2}$ and $C^*_{gf} = T_{gf}\left\{V(T_{gf})\right\}^{-1/2}$ under H_f and H_g, respectively, are approximately standard normal variates, two-tailed tests can be performed. For example, if C^*_{fg} is significantly negative, there is evidence of a departure from H_f in the direction of H_g. If C^*_{fg} is significantly positive, there is evidence of a departure from H_f in the direction opposite to H_g. The possible outcomes when both tests are performed are shown in Table 2.1. The decision-related terms “accept” and “reject” are used for simplicity. Rejection of both hypotheses suggests that it is necessary to look elsewhere for an appropriate model. Acceptance of both implies that there is no evidence that allows one to choose between the two models. Possible acceptance suggests that further testing is required, because although one model is not rejected, the other is rejected in favor of alternatives in a direction opposite to that of the model that is not rejected.

Example 2.2 (Example 2.1 cont.) Under H_L from (2.3), the estimator $\hat{\beta}$ converges in probability to $\beta_\alpha = \exp(\alpha_1 + \alpha_2/2)$, that is, β_α is the mean of the lognormal distribution. Further, expressions (2.5) and (2.6) become

Table 2.1 Possible outcomes of hypothesis tests for a pair of separate families

C_{gf}	C_{fg}		
	Significantly negative	Not significant	Significantly positive
Significantly negative	Reject both	Accept H_f	Reject both
Not significant	Accept H_g	Accept both	Possible acceptance of H_g
Significantly positive	Reject both	Possible acceptance of H_f	Reject both

$$T_{LE} = n\log(\hat{\beta}/\hat{\beta}_{\hat{\alpha}}), \qquad V_L(T_{LE}) = n\left(e^{\alpha_2} - 1 - \alpha_2 - \frac{\alpha_2^2}{2}\right), \tag{2.9}$$

where $\beta_{\hat{\alpha}} = \exp(\hat{\alpha}_1 + \hat{\alpha}_2/2)$.

Suppose that H_L and H_E change roles, such that the null distribution is exponential and the alternative is lognormal. Under H_E, from (2.3), the estimators $\hat{\alpha}_1$ and $\hat{\alpha}_2$ converge in probability to $\alpha_{1\beta} = \psi(1) + \ln\beta$ and $\alpha_{2\beta} = \psi'(1)$, respectively, that is, $\alpha_{1\beta}$ and $\alpha_{2\beta}$ are the mean and variance of the logarithm of a random variable with an exponential distribution, where $\psi(x) = d\ln\Gamma(x)/dx$, etc. For H_E, we asymptotically obtain

$$T_{EL} = n\left(\hat{\alpha}_1 - \alpha_{1\hat{\beta}} + 1/2\ln(\hat{\alpha}_2/\alpha_{2\hat{\beta}})\right), \tag{2.10}$$
$$V_E(T_{EL}) = n\left\{\psi'(1) - 1/2 + \psi''(1)/\psi'(1) + \psi'''(1)/4\{\psi'(1)\}^2\right\} = 0.2834n.$$

Example 2.3 (Pereira 1978, 1979) The hypotheses considered are that the distributions are lognormal, Weibull or gamma in nature:

$$H_L : f_L(y,\alpha_1,\alpha_2) = y(2\pi\alpha_2)^{-1/2}\exp\left\{-(\log y - \alpha_1)^2/2\alpha_2\right\},\ \alpha = (\alpha_1,\alpha_2),$$
$$H_W : f_W(y,\beta_1,\beta_2) = \beta_2/y\,(y/\beta_1)^{\beta_2}\exp\left\{-(y/\beta_1)^{\beta_2}\right\},\ \beta = (\beta_1,\beta_2), \tag{2.11}$$
$$H_G : f_G(y,\gamma_1,\gamma_2) = (y/\gamma_1)^{\gamma_2}/y\Gamma(\gamma_2)\exp\{-y_2/\gamma_1\},\ \gamma = (\gamma_1,\gamma_2).$$

(i) First, suppose that the null hypothesis is H_L and that the alternative is H_W. From (2.2), (2.3) and (2.5), we obtain, respectively,

$$\beta_{1\alpha} = \exp\{\alpha_1 + \sqrt{\alpha_2}/2\}, \quad \beta_{2\alpha} = \alpha_2^{-1/2},$$

$$T_{LW} = n\left\{\hat{\beta}_2\ln\hat{\beta}_1 - \beta_{2\hat{\alpha}}\ln\beta_{1\hat{\alpha}} - \ln\hat{\beta}_2 + \ln\beta_{2\hat{\alpha}} - \hat{\alpha}_1(\hat{\beta}_2 - \beta_{2\hat{\alpha}})\right\}, \tag{2.12}$$

$$V_L(T_{LW}) = 0.2183n.$$

When H_L and H_W change roles, such that the null hypothesis is H_W and the alternative is H_L, we have

$$\alpha_{1\hat{\beta}} = -0.5772/\beta_2 + \log\beta_1; \quad \alpha_{2\beta} = 1.6449/\beta_2^2,$$

$$T_{WL} = n\left\{\hat{\beta}_2(\hat{\alpha}_1 - \alpha_{1\hat{\beta}}) + \tfrac{1}{2}\ln(\hat{\alpha}_2/\alpha_{2\hat{\beta}})\right\}, \tag{2.13}$$

$$V_W(T_{WL}) = 0.2834n.$$

(ii) Suppose that the null hypothesis is H_L and the alternative is H_G; then, we have

$$\gamma_{1\alpha} = \exp\{\alpha_1 + \alpha_2/2\}, \ \ln\gamma_{2\alpha} - \psi(\gamma_{2\alpha}) = \ln\gamma_{1\alpha} - \alpha_1 = \alpha_2/2,$$
$$T_{LG} = n\left\{\ln\Gamma(\hat{\gamma}_2) - \hat{\gamma}_2\Gamma(\hat{\gamma}_2) + \hat{\gamma}_2 - \ln\Gamma(\gamma_{2\hat{\alpha}}) - \gamma_{2\hat{\alpha}}\psi(\gamma_{2\hat{\alpha}}) - \gamma_{2\hat{\alpha}}\right\}, \tag{2.14}$$
$$V_L(T_{LG}) = n\gamma_{2\alpha}^2\left[\exp(\alpha_2) - 1 - \alpha_2 - \frac{\alpha_2^2}{2}\right],$$

where $\gamma_{2\hat{\alpha}}$ is unique.

When H_L and H_G change roles, such that the null hypothesis is H_G and the alternative is H_L, we have

$$\alpha_{1\gamma} = \psi(\gamma_2) - \ln(\gamma_2/\gamma_1), \quad \alpha_{2\gamma} = \psi'(\gamma_2),$$

$$T_{GL} = \tfrac{n}{2}\ln(\hat{\alpha}_2/\alpha_2\hat{\gamma}), \tag{2.15}$$

$$V_G(T_{GL}) = n\left\{\frac{\psi'''(\gamma_2)}{4\{\psi'(\gamma_2)\}^2} - \frac{\gamma_2\{\psi''(\gamma_2)\}^2}{4\{\psi'(\gamma_2)\}^2\{\gamma_2\psi'(\gamma_2)-1\}} + 1/2\right\}.$$

(iii) Finally, consider the case in which the null hypothesis H_G is the gamma distribution and the alternative H_W is the Weibull distribution. Note that H_G and H_W are partially nonnested because for $\beta_2 = \gamma_2 = 1$, we specify the exponential distribution.
In this case,

$$\psi(\beta_{2\hat{\gamma}} + \gamma_2) - \frac{1}{\beta_{2\hat{\gamma}}} = \psi(\gamma_2), \ \ \ln\beta_{1\gamma} = \ln\left(\frac{\gamma_1}{\gamma_2}\right) + \beta_{2\gamma}^{-1}\ln\frac{\Gamma(\beta_{2\gamma}+\gamma_2)}{\Gamma(\gamma_2)},$$

$$T_{GW} = n\Bigg[\ln\left(\frac{\beta_{2\hat{\gamma}}}{\hat{\beta}_2}\right) - (\beta_{2\hat{\gamma}}\ln\beta_{1\hat{\gamma}} - \hat{\beta}_2\ln\hat{\beta}_1)$$
$$+\{\beta_{2\hat{\gamma}} - \hat{\beta}_2\}\left\{\psi(\hat{\gamma}_2) - \ln\left(\frac{\hat{\gamma}_2}{\hat{\gamma}_1}\right)\right\}\Bigg], \tag{2.16}$$

$$V_G(T_{GW}) = n\Bigg[\frac{\Gamma(2\beta_{2\gamma}+\gamma_2)\Gamma(\gamma_2)}{\{\Gamma(\beta_{2\gamma}+\gamma_2)\}^2}$$
$$+\frac{1}{\{\gamma_2\psi'(\gamma_2)-1\}\beta_{2\gamma}^2}\left\{3\beta_{2\gamma}^2 - \gamma_2 - \beta_{2\gamma}^4\psi'(\gamma_2) - \gamma_2\psi'(\gamma_2)\beta_{2\gamma}^2\right\}\Bigg],$$

where $\beta_{1\gamma}$ and $\beta_{2\gamma}$ are unique. When H_W is the null hypothesis and H_G is the alternative, we have

$$\gamma_{1\beta} = \beta_1 \Gamma\left(1 + \frac{1}{\beta_2}\right), \;\; \ln \gamma_{2\beta} - \psi(\gamma_{2\beta}) = \ln \Gamma\left(1 + \frac{1}{\beta_2}\right) - \frac{\psi(1)}{\beta_2},$$
$$T_{WG} = n\Big[\hat{\gamma}_{2\beta}\left\{\psi(\gamma_{2\hat{\beta}}) - 1\right\} - \ln \Gamma(\gamma_{2\hat{\beta}}) - \hat{\beta}_2\left\{\psi(\gamma_{2\hat{\beta}}) - \ln\left(\frac{\gamma_{2\hat{\beta}}}{\hat{\gamma}_1}\right)\right\}$$
$$- \left(\hat{\gamma}_2\left\{\psi(\hat{\gamma}_2) - 1\right\} - \ln \Gamma(\hat{\gamma}_2) - \hat{\beta}_2\left\{\psi(\hat{\gamma}_2) - \ln\left(\frac{\hat{\gamma}_2}{\hat{\gamma}_1}\right)\right\}\right)\Big],$$
$$V_W(T_{WG}) = n\Bigg[\left(\frac{\beta_2 - \gamma_{2\beta}}{\beta_2}\right)^2 \psi'(1) + \gamma_{2\beta}^2 \frac{\Gamma\left(1+\frac{2}{\beta_2}\right)}{\left\{\Gamma\left(1+\frac{1}{\beta_2}\right)\right\}^2} - \gamma_{2\beta}^2 - 1 \tag{2.17}$$
$$+2\left(\gamma_{2\beta} - \frac{\gamma_{2\beta}}{\beta_2}\right)\left\{\psi\left(1 + \frac{1}{\beta_2}\right) - \psi(1)\right\}$$
$$- \frac{1}{\psi'(1)}\left\{1 - \frac{\gamma_{2\beta}}{\beta_2}\left\{\psi\left(1 + \frac{1}{\beta_2}\right) - \psi(1)\right\}\right\}^2\Bigg].$$

Example 2.4 (Pereira 1978) We consider the model defined by

$$\log y_i = \mu + \sum_{r=1}^{m} z_{ir}\theta_r + \log u_i, \tag{2.18}$$

where the z_{ir} are the fixed values of the m regressors, μ is the unknown general mean, the θ_r are the unknown regression coefficients, and the u_i are iid random variables with density $f(u, \lambda)$, where f is a specified function and λ is an unknown scale or shape parameter.

As usual, it is assumed without loss of generality that

$$\sum_{i=1}^{n} z_{ir} = 0 \;\; (r = 1, \ldots, m). \tag{2.19}$$

It is also assumed, to permit the application of asymptotic theory, that if $z_i = (z_{i1}, \ldots, z_{im})$ and Z is an $n \times m$ matrix with rows z_i, then

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} z_i' z_i = \lim_{n\to\infty} n^{-1} Z'Z \text{ is a bounded positive definite matrix.}$$

Four particular cases are considered, defined by the form of the density of u_i as follows:

(a) Hypothesis H_L, a lognormal regression model, where $\log u_i$ is distributed as $N(0, \lambda)$.
(b) Hypothesis H_W, a Weibull regression model, where u_i is distributed in standard Weibull form with parameter λ, that is, with density $\lambda v^{\lambda-1} \exp(-v^{\lambda})$, equivalent to $v = v^{\lambda}$ with the standard exponential distribution with density e^{-v}.
(c) Hypothesis H_G, a gamma regression model, where u_i is distributed as $\lambda^{-1}G(\lambda)$; here, $G(\lambda)$ denotes a random variable with the standard gamma distribution shape parameter λ, with density $v^{\lambda-1}e^{-v}/\Gamma(\lambda)$.
(d) Hypothesis H_E, an exponential regression model, where u_i has a standard exponential distribution; this is a special case of (b) and (c) with $\lambda = 1$.

For comparisons of two hypotheses, different symbols are required for the set of unknown parameters $\left\{\mu, \lambda, \theta' = (\theta_1, \ldots, \theta_m)\right\}$ (omitting λ in (d)). This set will be denoted by $\alpha = \left(\alpha_1, \alpha_2, a'\right)$ for H_L, $\beta = \left(\beta_1, \beta_2, b'\right)$ for H_W, $\gamma = \left(\gamma_1, \gamma_2, c'\right)$ for H_G and $\boldsymbol{\beta} = \left(\beta, d'\right)$ for H_E. If (2.18) is assumed, then the information matrices $I(\mu, \lambda, \theta)$ for the models are block diagonal with blocks $I(\mu, \lambda)$ and $I(\theta)$ (Cox and Hinkley 1978).

The following results are obtained (Pereira 1978):

1. The estimates of the regression coefficients always converge to the true regression coefficients. For example, if H_L is the null hypothesis and H_W is the alternative, then the estimator $\hat{b} \xrightarrow{p} b_\alpha = a$. Section 2.5 investigates $V_\alpha(\hat{b})$ compared with $V_\alpha(\hat{a})$.
2. For all tests, the final expressions for the Cox test are equal to those from (2.9) through (2.16), presented in Examples 2.1 and 2.2. In these cases, the limits in probability are as follows (see Pereira 1978).

True model L, false model W:

$$\beta_{1\alpha} = \alpha_1 + \left(\frac{\alpha_2}{2}\right)^{-\frac{1}{2}}, \quad \beta_{2\alpha} = \left(\frac{1}{\sqrt{\alpha_2}}\right), \quad b_L = a.$$

True model W, false model L:

$$\alpha_{1\beta} = \beta_1 + \frac{\psi(1)}{\beta_2}, \quad \alpha_{2\beta} = \frac{\psi'(1)}{\beta_2^2}, \quad a_W = b.$$

True model L, false model G:

$$\gamma_{1\alpha} = \alpha_1 + \frac{\alpha_2}{2}, \quad \ln \gamma_{2\alpha} - \psi(\gamma_{2\alpha}) = \frac{\alpha_2}{2}, \quad d_L = a.$$

True model G, false model L:

$$\alpha_{1\gamma} = \psi(\gamma_2) - \log \gamma_2 - \gamma_1, \quad \alpha_{2G} = \psi'(\gamma_2), \quad a_G = c$$

True model G, false model W:

$$\beta_{1\gamma} = \gamma_1 - \ln \gamma_2 + \beta_{2\gamma}^{-1} \ln \left\{ \frac{\Gamma(\beta_{2\gamma})}{\Gamma(\gamma_2)} \right\}, \quad \psi(\beta_{2\gamma} + \gamma_2) - \beta_{2\gamma}^{-1} = \psi(\gamma_2), \; b_G = c.$$

True model W, false model G:

$$\gamma_{1\beta} = \beta_1 + \ln \Gamma\left(1 + \frac{1}{\beta_2}\right), \; \ln \gamma_{2\beta} - \psi(\gamma_{2\beta}) = \ln \Gamma\left(1 + \frac{1}{\beta_2}\right) - \frac{\psi(1)}{\beta_2}, \; c_w = b.$$

Example 2.5 (Example 1.4 cont.) Rewriting the models from Example 1.4 in matrix notation, we obtain

$$\begin{aligned} H_f &: y = X\alpha + u_f, \\ H_g &: y = Z\beta + u_g. \end{aligned} \tag{2.20}$$

Pesaran (1974) considered the case in which $u_f \sim N(0, \sigma_f^2 I_n)$ and $u_g \sim N(0, \sigma_g^2 I_n)$. Assuming that $\lim_{n\to\infty} \frac{1}{n} X'X = \sum_{x'x}$, $\lim_{n\to\infty} \frac{1}{n} Z'Z = \sum_{z'z}$ and $\lim_{n\to\infty} \frac{1}{n} X'Z = \sum_{x'z}$ exist and are finite, that $\sum_{x'x}$ and $\sum_{z'z}$ are non-singular and that $\sum_{x'z} \neq 0$, the Cox test of H_f against H_g is

$$\begin{aligned} \beta_\alpha &= \left(Z'Z\right)^{-1}(Z'X)\alpha, \quad \sigma_{g\alpha}^2 = \sigma_f^2 + \alpha' X' M_z X_\alpha, \\ T_{fg} &= \tfrac{n}{2} \ln \frac{\hat{\sigma}_f^2}{\sigma_{g\hat{\alpha}}}, \\ V(T_{fg}) &= \frac{\hat{\sigma}_f^2}{\sigma_{g\hat{\alpha}}^4} \hat{\alpha}' X' M_z M_x M_z X_{\hat{\alpha}}, \end{aligned} \tag{2.21}$$

where $M_x = I - X'(X'X)^{-1}X$ and $M_z = I - Z'(Z'Z)^{-1}$ (Pesaran 1974).

Pesaran and Deaton (1978) extended this test to nonlinear systems of equations, and Timm and Al-Subaihi (2001) also extended it to seemingly unrelated regression models. Araujo et al. (2005) extended it to systems of linear equations.

Pereira (1984) derived the Cox test for a similar problem

$$\begin{aligned} H_f &: \ln y_i = \alpha_1 + xa + \ln u_f, \\ H_g &: \ln y_i = \beta_1 + zb + \ln u_g, \end{aligned} \tag{2.22}$$

where $u_f \sim W(\alpha_2)$ and $u_g \sim W(\beta_2)$ are random variables with standard Weibull distributions with parameters α_2 and β_2, respectively.

Under the same assumptions regarding $X'X$, XZ, and $Z'Z$, we have

$$\begin{aligned} \beta_{1\alpha} &= \alpha_1 + \tfrac{1}{\beta_{2\alpha}} \ln \Gamma(K), \beta_{2\alpha} = [\psi(k) - \psi(1)]^{-1}\alpha_2, b_a = (Z'Z)^{-1}(Z'X)a, \\ T_{fg} &= (\hat{\alpha}_2 - \beta_{2\hat{\alpha}}) \sum_{i=1}^{n} \ln y_i + n(\hat{\alpha}_2 - \beta_{2\hat{\alpha}}) \left\{ \hat{\alpha}_1 + \frac{\psi'(1)}{\hat{\alpha}_2} \right\} \\ &\quad - n \ln \left(\frac{\hat{\beta}_2}{\beta_{2\hat{\alpha}}} \right) + n(\hat{\beta}_1 \hat{\beta}_2 - \beta_{1\hat{\alpha}} \beta_{2\hat{\alpha}}), \end{aligned} \tag{2.23}$$

where $k = 1 + \beta_{2\alpha}/\alpha_2$. Finally, an interesting result follows if the hypotheses in (2.22) are any two distributions (a) to (d) from Example 2.3. For instance, if u_f has an exponential distribution and u_g has a lognormal one, then the tests will have the same expressions as in (2.9) through (2.17) plus an additional equation corresponding to the limit in probability. In this case, this term is $b_L = (Z'Z)^{-1}Z^{-1}a$, and there are analogous terms for the other cases (Pereira 1978, 1981b, 1984).

Example 2.6 (Pereira 1981b) The results of Example 2.3 were applied to survival data for 93 malignant tumor patients collected in the Brain Tumour Study conducted by M.D. Anderson Hospital and the Tumour Institute. The complete description of the data set is presented in Pereira (1976). All patients received surgery and were randomized with regard to whether they received a chemotherapeutic agent (Mithramycin) or conventional care (Control) during the recovery period. The tumors were classified by their principal position in the brain. The other variables recorded were age, duration of symptoms (headache, personality change, motor deficit, etc.), sex, and level of radiation (see Walker et al. 1969). For each patient, a vector of covariates $\underline{z} = (z_1, \ldots, z_{10})$ was defined, where z_1, z_2, z_3, z_4 and z_5 represented age, duration of symptoms, sex, treatment and radiation, respectively. The remaining variates z_6, z_7, z_8, z_9, and z_{10} were indicators of the positions of the cancer cells, with one variate corresponding to each of the frontal, temporal, parietal, and occipital lobes and the deep BG/T region.

In the search for a suitable model, the simplest models were examined first. The exponential and lognormal regression models yielded statistic values of $T_{LE} = -2.813$, indicating a departure from H_L in the direction of H_E, and $T_{EL} = -2.909$, indicating a departure from H_E in the direction of H_L. This suggests that neither model fits the data well. Subsequently, departures from H_E in the directions of H_G and H_W were tested. Because these hypotheses are not separate, asymptotic normal distributions of the maximum likelihood estimators of the shape parameters of the gamma and Weibull regression models were used, or, equivalently, the asymptotic χ^2 distribution of the maximum likelihood ratio. The results are summarized in Table 2.2 and show that the null hypothesis of an exponential regression model is rejected under the assumption of either a Weibull or a gamma model. Note that the null hypothesis H_E is rejected more strongly by the Weibull test.

Next, H_L was tested against H_G and H_W. All test results and other values of interest are shown in Table 2.3. The test statistic $T_{LG} = -3.119$ rejects H_L in favor of H_G, and the test statistic $T_{LG} = -1.016$ suggests reasonable agreement with H_G. For H_W, the results were $T_{LW} = -3.699$, rejecting H_L, and $T_{WL} = 0.137$, suggesting good agreement with H_W. Again, H_L is rejected more strongly when compared with H_W.

Given the results of the tests carried out above, the remaining two possible working hypotheses are H_G and H_W. As seen in Table 2.3, the test statistic $T_{GW} = -2.436$ points to a departure from H_G in the direction of H_W, and the test statistic $T_{WG} =$

Table 2.2 Testing for an exponential regression model

Alternative	MLE		Likelihood ratio	
	Normal deviate	Significance level	$-2\log\lambda$	Significance level
Gamma	3.982	0.000035	26.765	<0.00001
Weibull	5.084	<0.00001	31.367	<0.00001

Table 2.3 Results of all tests of separate families of hypotheses

Test	Normal deviate	Significance level	Estimates of probability limits
T_{LE}	−2.813	0.00248	$\hat{\delta}_{1L} = 5.196$
T_{EL}	−2.909	0.00191	$\hat{\alpha}_{1E} = 4.557$, $\hat{\alpha}_{2E} = 1.645$
T_{LG}	−3.119	0.00090	$\hat{\gamma}_{1L} = 5.196$, $\hat{\gamma}_{2L} = 1.777$
T_{GL}	−1.016	0.15386	$\hat{\alpha}_{1G} = 4.890$, $\hat{\alpha}_{2G} = 0.533$
T_{LW}	−3.699	0.00011	$\hat{\beta}_{1L} = 5.281$, $\hat{\beta}_{2L} = 1.277$
T_{WL}	0.137	0.44433	$\hat{\alpha}_{1W} = 4.906$, $\hat{\alpha}_{2W} = 0.570$
T_{GW}	−2.436	0.00734	$\hat{\beta}_{1G} = 5.244$, $\hat{\beta}_{2G} = 1.560$
T_{WG}	0.967	0.16602	$\hat{\gamma}_{1W} = 5.132$, $\hat{\gamma}_{2W} = 2.367$
$\hat{\alpha}_1 = 4.8896$ $\hat{\beta}_1 = 5.2461$ $\hat{\gamma}_1 = \hat{\delta}_1 = 5.1338$			
$\hat{\alpha}_2 = 0.6137$ $\hat{\beta}_2 = 1.6989$ $\hat{\gamma}_2 = 2.1999$			

0.967 suggests good agreement of the hypothesis H_W with these data. Therefore, the Weibull regression model should be used for further analysis of the data.

The models can thus be ranked in order of preference as dictated by test results as follows: the Weibull regression model is ranked first, followed by the gamma, lognormal and exponential regression models. This is also the ordering indicated by the maxima of the log-likelihood functions, which are $\hat{\ell}_W = -554.81$, $\hat{\ell}_G = -557.06$, $\hat{\ell}_L = -563.94$ and $\hat{\ell}_E = -570.44$.

Finally, the results obtained for Example 2.3 show that all estimators of the regression coefficients are consistent, independent of distributional assumptions. Therefore, the efficiencies of the estimators of the regression coefficients when an incorrect model is used compared with the case of the correct model can be investigated. It will be shown in Sect. 2.5 that when the correct model is a Weibull regression model with $\beta_2 = 1.669$, these efficiencies are 0.61 for the lognormal regression model and 0.95 for the gamma and exponential regression models.

2.3 A Test Based on a Compound Model

Silva (2001) embedded the models specified by $H_f : f(y, \alpha)$ and $H_g : g(y, \beta)$ in the general model

$$h_c(y, \rho, \lambda, \alpha, \beta) = \frac{[\lambda f^{\rho}(y, \alpha) + (1 - \lambda) g^{\rho}(y, \beta)]^{\frac{1}{\rho}}}{\int [\lambda f^{\rho}(y, \alpha) + (1 - \lambda) g^{\rho}(y, \beta)]^{\frac{1}{\rho}} \, dy}. \tag{2.24}$$

If $\rho = 1$, Eq. (2.24) becomes

$$h_l(y, \lambda, \alpha, \beta) = \lambda f_1(y, \alpha) + (1 - \lambda) g(y, \beta). \tag{2.25}$$

Taking the limit as $\rho \to 0$, Eq. (2.24) becomes

$$h_e(y, \lambda, \alpha, \beta) = \frac{f_1^{\lambda}(y, \alpha) g^{1-\lambda}(y, \beta)}{\int f_1^{\lambda}(y, \alpha) g^{1-\lambda}(y, \beta) dy}. \tag{2.26}$$

Silva obtained the Rao score function for the general distribution (2.24).

Thus far, we have been interested in testing H_f against H_g. Now, we will address expressions (2.25) and (2.26), in turn.

Let us first consider the test of $\lambda = 1$ in Cox's exponential compound model (2.26), developed by Atkinson (1970).

It has been shown (Pesaran 1981; Antle and Bain 1969; Silva 2001) that a test of $\lambda = 1$ can be obtained using the Rao score test procedure.

The log-likelihood function of the compound model is

$$\ell(\lambda, \alpha, \beta) = \lambda \ell_f(\alpha) + (1 - \alpha) \ell_g(\beta) - \int \left[\ell_f^{\lambda}(\alpha) \ell_g^{1-\lambda}(\beta) \right] dy. \tag{2.27}$$

Two possible tests can be considered (Pesaran 1981):

(i) The parameters α_0 and β_0 are known.
In this case,

$$\frac{\partial}{\partial \lambda} \ell_{\lambda}(\lambda) = \ell_f(\alpha_0) - \ell_g(\beta_0) - E_f \left\{ \ell_f(\alpha_0) - \ell_g(\beta_0) \right\} = \ell_{fg} - E(\ell_{fg}). \tag{2.28}$$

Therefore, the Rao score test statistic is

$$RS(\alpha_0, \beta_0) = \frac{\left(\ell_{fg} - E(\ell_{fg}) \right)^2}{V(\ell_{fg})}, \tag{2.29}$$

and this is related to the statistics discussed in Atkinson (1969, 1970) for choosing among prediction formulas.

(ii) The parameters α_0 and β_0 are known.
Under the null hypothesis $H_f : \lambda = 1$, the information matrix corresponding to the parameters (λ, α, β) is singular because β is non-identifiable. Adding the information provided by the null hypothesis and working with a smaller order information matrix, the log-likelihood function becomes (Dastoor 1985)

$$\ell_{\lambda}(\lambda, \alpha) = \lambda \ell_f(\alpha) - (1 - \lambda) \ell_f(\hat{\beta}) - \log \left\{ \int f^{\lambda}(y, \beta) g^{1-\lambda}(y, \hat{\beta}) dy \right\}, \tag{2.30}$$

because $\hat{\beta} \longrightarrow \beta_\alpha$, by assumption. Consequently, the score vector is

$$\begin{bmatrix} \frac{\partial \ell_\lambda}{\partial \lambda}(\lambda, \alpha) \\ \\ \frac{\partial \ell_\lambda}{\partial \alpha}(\lambda, \alpha) \end{bmatrix} = \begin{bmatrix} \ell_0(\hat{\alpha}) - \ell_1(\hat{\beta}) - E_\alpha \left[\ell_0(\alpha) - \ell_1(\hat{\beta}) \right]_{\alpha=\hat{\alpha}} \\ \\ 0 \end{bmatrix}$$
$$\approx \begin{bmatrix} \ell_0(c) - \ell_1(\hat{\beta}) - E_\alpha \left[\ell_0(\alpha) - \ell_1(\beta) \right]_{\alpha=\alpha} \\ \\ 0 \end{bmatrix} \tag{2.31}$$
$$= \begin{bmatrix} T_{fg} \\ \\ 0 \end{bmatrix},$$

and the Rao score test statistic is

$$RS = \left[\frac{T_{fg}}{V(T_{fg})} \right]^2, \tag{2.32}$$

which is exactly the square of the Cox test statistic.
The Atkinson (1970) test statistic is obtained similarly by replacing $\hat{\beta}$ by $\beta_{\hat{\alpha}}$. Cox's and Atkinson's test statistics have the same estimated variance (Pereira 1977a).

We now consider the linear compound model (2.5). Using the linear compound model estimated based on the maximum likelihood, Quandt (1974) suggested the following procedures:

- Let $\hat{\lambda}$ and $\hat{\sigma}_\lambda^2$ denote the maximum likelihood estimate for λ and the asymptotic variance of $\hat{\lambda}$, respectively. The hypothesis H_f is rejected if the interval $(\hat{\lambda} - z_p\hat{\sigma}_\lambda, \hat{\lambda} + z_p\hat{\sigma}_\lambda)$ does not overlap 1.0 for z, which is the corresponding normal variate for a level of significance p. The hypothesis H_g is rejected if it does not overlap 0.0. Both hypotheses are rejected if the interval overlaps neither 1.0 nor 0.0. Finally, neither hypothesis should be rejected if the interval overlaps both 1.0 and 0.0.
- Consider the log-likelihood ratios RL_f and RL_g:

$$\begin{aligned} RL_f &= \ell_f(y, \hat{\alpha}) - \ell_l(\hat{\lambda}, \hat{\alpha}, \hat{\beta}), \\ RL_g &= \ell_g(y, \hat{\beta}) - \ell_l(\hat{\lambda}, \hat{\alpha}, \hat{\beta}). \end{aligned} \tag{2.33}$$

A large value of $(-2)\times$ either likelihood ratio leads to the rejection of the corresponding hypothesis.

The difficulties with regard to numerical methods that Quandt faced when obtaining the maximum likelihood estimates of the parameters and the corresponding asymptotic covariance matrix are greatly reduced by the Expectation–Maximization (EM) algorithm. For these results, see Oakes (1999) and Lanot (2002).

Example 2.7 (Pereira 1976, 1977a, 1981a) Consider the distributions and notations of Examples 2.1 and 2.2: the lognormal, Weibull, gamma and exponential distributions. The Atkinson test statistics for these cases are as follows:

(i)
$$T_{LE}(A) = n\left[\frac{\hat{\beta}}{\beta_{\hat{\alpha}}} - 1\right],$$

(ii)
$$T_{EL}(A) = n\left\{\hat{\alpha}_1 - \alpha_{1\hat{\beta}} + \frac{1}{2\alpha_{2\hat{\beta}}}\left[\hat{\alpha}_2 - \alpha_{2\hat{\beta}} + (\hat{\alpha}_1 - \alpha_{1\hat{\beta}})^2\right]\right\},$$

(iii)
$$T_{LW}(A) = \sum_{i=1}^{n}\left[\frac{y_i}{\beta_{1\hat{\alpha}}}\right]^{\beta_{2\hat{\alpha}}},$$

(iv)
$$T_{WL}(A) = n\left[\hat{\beta}_2(\hat{\alpha}_1 - \alpha_{1\hat{\beta}}) + \frac{1}{2\alpha_{2\hat{\beta}}}(\hat{\alpha}_2 - \alpha_{2\hat{\beta}}) + (\hat{\alpha}_1 - \alpha_{1\hat{\beta}})^2\right],$$

(v)
$$T_{LG}(A) = n\gamma_{2\hat{\alpha}}\left[\frac{\hat{\gamma}_1}{\gamma_{1\hat{\alpha}}} - 1\right],$$

(vi)
$$T_{GL}(A) = n\left[\frac{\hat{\alpha}_2}{\alpha_{2\hat{\gamma}}} - 1\right],$$

(vii)
$$T_{GW}(A) = \left\{\sum_{i=1}^{n}\left[\frac{y_i}{\beta_{1\hat{\gamma}}}\right]^{\beta_{2\hat{\gamma}}} - n\right\},$$

(viii)
$$T_{WG}(A) = n\left\{(\hat{\beta}_2 - \gamma_{2\hat{\beta}})\left[\psi(\hat{\gamma}_2) - \ln\left(\frac{\hat{\gamma}_2}{\hat{\gamma}_1}\right) - \psi(\gamma_{2\hat{\beta}})\right]\right\}$$
$$+n\left\{(\hat{\beta}_2 - \gamma_{2\hat{\beta}})\left[\ln\left(\frac{\gamma_{2\hat{\beta}}}{\gamma_{1\hat{\beta}}}\right) + (\hat{\gamma}_2 - \gamma_{2\hat{\beta}})\right]\right\}.$$

Example 2.8 (Example 2.5 cont.) Consider again the hypotheses $H_f : y = X\alpha + u_f$, where $u_f \sim N(0, \sigma_f^2 I_n)$, and $H_g : y = Z\beta + u_g$, where $u_g \sim N(0, \sigma_g^2 I_n)$. An exponential compound model that includes these two models, after integration and simplification, becomes

$$\begin{aligned} H_\lambda : y &= \left\{\lambda \frac{\sigma^2}{\sigma_f^2}\right\} y\alpha + \left\{(1-\lambda)\frac{\sigma^2}{\sigma_g^2}\right\} z\beta + u \\ &= \xi x\alpha + (1-\xi)z\beta + u \\ &= x\gamma_1 + z\gamma_2 + u, \end{aligned} \tag{2.34}$$

where $u \sim N(0, \sigma^2 I)$ and $\sigma^2 = \frac{\sigma_f^2 \sigma_g^2}{\{\lambda\sigma_g^2 + (1-\lambda)\sigma_f^2\}}$.

Now, let us return our attention to the problem of testing H_f against H_g by testing $\lambda = 1$.

Because γ_1 and γ_2 are estimable, we can choose between the models by examining their t-statistics, but we cannot identify $(\lambda, \alpha, \beta, \sigma_f^2, \sigma_g^2)$ separately.

Alternatively, the Rao score statistic can be applied as in (2.29) to overcome the difficulty of the nonexistence of β under H_f. During the 1980s, econometricians developed a number of practical alternatives by replacing the parameter β of the alternative hypothesis in (2.34) with some reasonable estimator, such as $\hat{\beta}$, the maximum likelihood estimator of β under H, or $\beta_{\hat{\alpha}}$, a consistent estimate of the probability limit of $\hat{\beta}$ defined in (2.22). In these cases, the resulting equations respectively become

$$\begin{aligned} y &= \xi x\alpha + (1-\xi)z\hat{\beta} + u, \\ y &= \xi x\alpha + (1-\xi)z\beta_{\hat{\alpha}} + u. \end{aligned} \tag{2.35}$$

The t-tests thus obtained are called the J test of Davidson and MacKinnon (1981, 1982) and the JA test of Fisher and McAleer (1981), respectively.

Alternative estimates for nonlinear extensions are discussed further in Pesaran (1982), Fisher (1983) and McAleer (1995).

Extensions to simultaneous equations are presented in Pesaran (1982) and Davidson and MacKinnon (1983), with divergent results related to their applicability. Pesaran suggests that only the Cox test can be extended to the multivariate case without unreasonable assumptions.

McAleer (1995) also presents a classificatory review of the empirical nonnested models and tests described in this example.

Example 2.9 (Quandt 1974) Three procedures were considered for testing alternative econometric equations: Pesaran's procedure developed for the Cox test (Pesaran 1974), Cox's exponential compound procedure (Atkinson 1970), and Cox's linear compound procedure (Quandt 1974). The hypotheses specified were

$$\begin{aligned} H_f &: y_t = \alpha_1 + \alpha_2 y_{t-1} + \alpha_3 M_{t-1} + \alpha_3 N_t + u_t, u_t \sim N(0, \sigma_u^2), \\ H_g &: y_t = \beta_1 p_t + \beta_2 y_{t-1} + \beta_3 M_{t-1} + \beta_4 N_t + v_t, v_t \sim N(0, \sigma_v^2), \end{aligned} \tag{2.36}$$

Table 2.4 Test procedures

Test results		Decision
Linear compound	$\hat{\lambda} = 800$ $\hat{\sigma}_\lambda = 0.088$	Reject both
$-2 \times$ likelihood ratio	$-2RL_f = 23.18$ $-2RL_g = 13.32$	Reject both
Cox	$C_{fg} = 0.106$ $C_{gf} = -1.077$	Accept both
Exponential compound	$t_f = 1.035$ $t_g = 0.107$	Accept both

where y_t is the per capita disposable income, M_t is the total per capita deposit and the currency outside of banks, I_t is the gross per capita investment, G_t is the per capita government expenditure on goods and services, p_t is the cost of living index, T_t is the per capita GNP minus y_t, and $N_t = (I_t + G_t - T_t)$. Quandt (1974) applied his procedure using data available in the literature. The results are presented in Table 2.4. They suggest that the consumption function is a hybrid of the equations expressed in H_f and H_g.

2.4 Alternative Tests

2.4.1 Test for Multiple Hypotheses

Sawyer (1984) introduced a statistic to test a currently held set of hypotheses against a series of M alternatives. This test avoids the problems that arise when several hypotheses are under consideration and binary comparisons of them are made, that is, comparisons of two hypotheses at a time.

Without loss of generality, we consider only three hypotheses: H_f, H_g, and H_h. The test relies on the results of the Cox test. Suppose that the null hypothesis is H_f, and consider the vector of Cox test statistics:

$$T_f' = (T_{fg}, T_{fh}). \tag{2.37}$$

T_f is asymptotically normally distributed as a bivariate (M-1=2) normal distribution with a vector mean of zero and a covariance of $\sum = \sigma_{ij}$, $j = g, h$, given by

$$\sigma_{ij} = Cov_f(T_{fg}, T_{fh}) - Cov_f\left(z\frac{\partial \ell_f(\alpha)}{\partial \alpha}\right)' I^{-1}(\alpha) Cov\left(z\frac{\partial \ell_f(\alpha)}{\partial \alpha}\right), \tag{2.38}$$

as obtained using expression (2.5).

The multiple test for testing H_f against the separate alternatives H_g and H_h is

$$T_f'\left(\sum\right)^{-1} T_f, \tag{2.39}$$

which is asymptotically distributed as a χ_2^2 (χ_{M-1}^2) random variable. Values that exceed the critical value indicate the rejection of H_f. The maximum likelihood estimators of α provide consistent estimator of $\sum$.

Sawyer showed that this test is equivalent to the Rao score (or Lagrange multiplier) test for $\lambda_f = 1$ and $\lambda_g = \lambda_h = 0$ in the exponential mixture model:

$$f(y, \alpha, \beta, \gamma) = k(\alpha, \beta, \gamma) f^{\lambda_f}(y, \alpha) g^{\lambda_g}(y, \beta) h^{\lambda_h}(y, \gamma), \tag{2.40}$$

where $\lambda_i' \geq 0$ and $\lambda_f + \lambda_g + \lambda_h = 1$. With ℓ_λ denoting the log-likelihood for this model and $\lambda' = (\lambda_g, \lambda_h)$, the resulting test statistic is

$$\left(\frac{\partial l_\lambda}{\partial \lambda}\right)' I''(\lambda) \left(\frac{\partial l_\lambda}{\partial \lambda}\right)$$

evaluated at $\lambda = 0$. $I''(\lambda)$ is the sub-matrix corresponding to λ in the information matrix corresponding to the model in (2.40).

Example 2.10 (Sawyer 1984) Consider the hypotheses

$$\begin{aligned} H_f &: \ f(y, \beta) = \beta^{-1} \exp\left(-\tfrac{y}{\beta}\right), \\ H_g &: \ g(y, \alpha_1, \alpha_2) = y^{-1} (2\pi_\alpha)^{-\frac{1}{2}} \exp\left\{-\tfrac{(\log y - \alpha_1)^2}{2\alpha_2}\right\}, \\ H_h &: \ h(y, p, \gamma) = y^{-1} \left(\tfrac{y}{\gamma}\right)^p \exp\left(-\tfrac{y}{\gamma}\right), \ p \neq 1 \text{ known.} \end{aligned} \tag{2.41}$$

The terms required for the test statistic are obtained from Cox (1961, p. 117)

$$\begin{aligned} T_{fg} &= (\hat{\alpha}_1 - \alpha_{1\hat{\beta}}) + \tfrac{1}{2} \log\left(\tfrac{\hat{\alpha}}{\psi'(1)}\right), \\ T_{fh} &= -(p-1)(\hat{\gamma} - \gamma_{\hat{\beta}}), \\ \sigma_{gg} &= V_E(T_{EL}) = \tfrac{0.2833}{n}, \\ \sigma_{hh} &= V_E\left(T_{EG_p}\right) = (p-1)^2 \tfrac{(\psi'(1)-1)}{n} = \tfrac{(p-1)^2 0.6449}{n}, \\ \sigma_{gh} &= C_E(T_{fg}, T_{fh}) = (p-1) \tfrac{\left\{1 - \psi'(1) - \frac{\psi''(1)}{2\psi'(1)}\right\}}{n} \\ &= \tfrac{(p-1)0.0858}{n}. \end{aligned} \tag{2.42}$$

The test statistic is

$$T_f = \begin{pmatrix} T_{fg} & T_{fh} \end{pmatrix} \begin{pmatrix} \sigma_{gg} & \sigma_{gh} \\ \sigma_{gh} & \sigma_{hh} \end{pmatrix}^{-1} \begin{pmatrix} T_{fg} \\ T_{fh} \end{pmatrix}. \tag{2.43}$$

For regression models, Davidson and MacKinnon (1981) recommend the use of the J test (and its alternatives) to test for the true hypothesis against several alternatives at once. To test H_1 ($y = x\alpha + u_f$) against $(M-1)$ alternative models ($y = z_j \beta_j + v_{g_j}$) using the J test, one simply estimates

$$y = \left(1 - \sum_{j=1, j\neq m}^{M}\right) x\alpha + \sum_{j=1}^{M-1} \gamma_j z_j \hat{\beta}_j + v \tag{2.44}$$

and performs a likelihood ratio test of the requirement that all $\gamma_j (j \neq m)$ are zero.

Hagemann (2012) used (2.44) to test the validity of a model m in the presence of several alternatives by means of a Wald test J_m for $H_m : \gamma_j = 0,\ j = 1, \ldots, M,\ j \neq m$. He then argued that "if one of the models under consideration is the correct model, then its J_m statistic has a χ^2_{M-1} distribution and the statistics of the other models diverge; if, instead, the correct model is not among the M models, then all statistics will diverge. Thus, only the model with the smallest J statistic can possibly be the correct model and we reject the hypothesis that the correct model is one of those M considered when the smallest J statistic is large." This motivates the following alternative MJ test to traditional sequential testing:

1. For each m ($m = 1, \ldots, M$), perform regression (2.44) and compute J_m; define

$$MJ = \min\{J_m, m = 1, \ldots, M\}. \tag{2.45}$$

2. Reject all models m ($m = 1, \ldots, M$) if $MJ > \chi_{1-\alpha,M-1}$, where $\chi_{1-\alpha,M-1}$ is the $1 - \alpha$ quantile of the χ^2_{M-1} distribution.

Hagemann (2012) not only proved these results but also noted that they can be extended to nonlinear models and models with heteroscedastic and autocorrelated errors. Additionally, the related tests (JA, Cox, Atkinson, etc.) can be extended in an analogous manner.

Example 2.11 (Cribari-Neto and Lucena 2015) These authors extended the results of Hagemann to beta regression with alternative nonlinear forms of the regressors.

The beta regression of Ferrari and Cribari-Neto (2004) considers a beta density:

$$h(y, \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{2.46}$$

where $y \in (0, 1)$, $\mu \in (0, 1)$ and $\phi > 0$. Thus, $E(y) = \mu$ and $Var(y) = \mu(1 - \mu)\big/(1 + \phi)$.

For a sample of independent beta variables with mean μ_t and precision ϕ, the beta regression considers a linear predictor η_t that is related to the mean μ_t through a link function

$$g(\mu_t) = \eta_t = \sum_{i=1}^{k} x_{ti}\beta_i = x_t\beta, \tag{2.47}$$

where β is a vector of unknown parameters and x_t is a vector of observations on k regressors.

Cribari-Neto and Lucena (2015) considered five nonnested models with different link functions in the submodels of the mean, namely, logit, probit, log-log, complementary log-log and Cauchy functions, and with a varying precision parameter ϕ_t.

The data used (32 observations) considered the yield, namely, the proportion of crude oil converted into gasoline after distillation and fractionation, as the variable of interest. Two explanatory variables were used:

- temp: X_{10}—the temperature in degrees Fahrenheit at which all of the gasoline vaporized, and
- batch: $(X_1, \ldots, X_9)$—a factor indicating ten different batches of conditions considered in the experiment.

The models were as follows:

$$\begin{aligned}
m_1 &: \log\left(\frac{\mu_t}{1-\mu_t}\right) = \eta_1(X_t\beta), \quad \log(\phi_t) = \gamma_0 + \gamma_1 X_{t,10}, \\
m_2 &: \Phi^{-1}(\mu_t) = \eta_2(X_t\beta), \quad \phi_t = \phi, \\
m_3 &: -\log\{-\log(\mu_t)\} = \eta_3(X_t\beta), \quad \phi_t = \phi, \\
m_4 &: -\log\{-\log(1-\mu_t)\} = \eta_4(X_t\beta), \quad \log(\phi_t) = \gamma_0 + \gamma_1 X_{t,10}, \\
m_5 &: \tan\{\phi_t(\mu_t - 0.5)\} = \eta_5(X_t\beta), \quad \log(\phi_t) = \gamma_0 + \gamma_1 X_{t,10},
\end{aligned} \tag{2.48}$$

where $L(X_t\beta) = \beta_0 + \sum_{j=1}^{10} \beta_j X_{tj}$.

The J and MJ tests and the likelihood ratio and Wald tests were performed. The J test p-values for each pair of nonnested models are reported in Table 2.5.

The MJ p-values were 0.0094 and 0.0023 (LR and Wald statistics, respectively). Therefore, the authors concluded that the correct model was among the candidate models. Because the smallest J statistic was that of the log-log model, this model was selected.

The authors also confirmed this choice using other statistics and criteria and also presented Monte Carlo simulations of these tests and their bootstrap versions.

2.4.2 Test Based on Nondirectional Divergence

Consider the directed divergence known as the Kullback–Leibler information criterion (KLIC):

$$\begin{aligned}
&I_\alpha(f,g) = \int \left\{\ell_f(\alpha) - \ell_g(\beta)\right\} f(y,\alpha)dy = \tfrac{1}{n} E_\alpha \left\{\ell_f(\alpha) - \ell_g(\beta)\right\} \\
&\text{and} \\
&I_\beta(g,f) = \int \left\{\ell_g(\beta) - \ell_f(\alpha)\right\} g(y,\beta)dy = \tfrac{1}{n} E_\beta \left\{\ell_g(\beta) - \ell_f(\alpha)\right\}.
\end{aligned} \tag{2.49}$$

A nondirectional divergence between H_f and H_g is given by

Table 2.5 J test p-values obtained using the LR and Wald statistics for the five competing models

Model	LR	Wald
Logit versus probit	1.715×10^{-5}	2.637×10^{-8}
Logit versus log-log	1.828×10^{-5}	2.657×10^{-8}
Logit versus compl. log-log	0.0004	1.667×10^{-5}
Logit versus Cauchit	0.0023	0.0003
Probit versus logit	0.0016	0.0007
Probit versus log-log	0.0040	0.0001
Probit versus compl. log-log	0.0026	0.0013
Probit versus Cauchit	0.0089	0.0061
Log-log versus logit	0.4869	0.4863
Log-log versus probit	0.2634	0.2596
Log-log versus compl. log-log	0.5505	0.5501
Log-log versus Cauchit	0.7583	0.7584
Compl. log-log versus logit	1.629×10^{-5}	8.207×10^{-9}
Compl. log-log versus probit	8.581×10^{-7}	3.25×10^{-12}
Compl. log-log versuslog-log	1.496×10^{-6}	9.013×10^{-12}
Compl. log-log versusCauchit	0.0030	6.319×10^{-6}
Cauchit versus logit	5.4×10^{-8}	2.028×10^{-12}
Cauchit versus probit	6.01×10^{-9}	2.527×10^{-15}
Cauchit versus log-log	1.6×10^{-10}	$<2.2 \times 10^{-16}$
Cauchit versus compl. log-log	2.193×10^{-7}	6.624×10^{-11}

$$\begin{aligned} J(f, g) &= \int \{f(y, \alpha) - g(y, \beta)\} \ln \tfrac{f(y,\alpha)}{g(y,\beta)} \mathrm{d}y \\ &= I_\alpha(f, g) - I_\beta(g, f). \end{aligned} \tag{2.50}$$

This can be estimated by

$$\begin{aligned} \hat{J}(f, g) &= \hat{I}_{\hat{\alpha}}(f, g) + \hat{I}_{\hat{\beta}}(g, f) \\ &= \tfrac{1}{n}\left[E_{\hat{\alpha}}\left\{\ell_f(\hat{\alpha}) - \ell_g(\hat{\beta})\right\}\right] + \tfrac{1}{n}\left[E_{\hat{\beta}}\left\{\ell_g(\hat{\beta}) - \ell_f(\hat{\alpha})\right\}\right] \\ &= \tfrac{1}{n}\left[\mathrm{p}\lim_{n\to\infty}\{\ell_f(\hat{\alpha}) - \ell_g(\beta_{\hat{\alpha}})\}\right]_{\hat{\alpha}} + \tfrac{1}{n}\left[\mathrm{p}\lim_{n\to\infty}\left\{\ell_g(\hat{\beta}) - \ell_f(\alpha_{\hat{\beta}})\right\}\right]_{\hat{\beta}}. \end{aligned} \tag{2.51}$$

Sawyer (1983) proposed the following asymmetric test statistic for testing two separate hypotheses, H_f against H_g:

$$\begin{aligned} S_f(\hat{\alpha}) &= E_{\hat{\beta}}\left\{\ell_f(\hat{\alpha}) - \ell_g(\hat{\beta})\right\} - E_{\hat{\alpha}}\left[E_{\hat{\beta}}\left\{\ell_f(\hat{\alpha}) - \ell_g(\hat{\beta})\right\}\right] \\ &= n\left[I_{\hat{\beta}}(f, g) - E_{\hat{\alpha}} I_{\hat{\beta}}(f, g)\right]. \end{aligned} \tag{2.52}$$

Under H_f, $S_f(\hat{\alpha})$ has a mean of zero, and under H_f, it has a negative mean.

The analogous statistic

$$S_g(\hat{\beta}) = E_{\hat{\alpha}}\left\{\ell_g(\hat{\beta}) - \ell_f(\hat{\alpha})\right\} - E_{\hat{\beta}}\left[E_{\hat{\alpha}}\left\{\ell_g(\hat{\beta}) - \ell_f(\hat{\alpha})\right\}\right]$$
$$= n\left[I_{\hat{\alpha}}(g, f) - E_{\hat{\beta}} I_{\hat{\alpha}}(g, f)\right] \qquad (2.53)$$

is used to detect departures from the null hypothesis H_g against the alternative H_f.

Sawyer also showed that

$$V_\alpha\left\{S_f(\hat{\alpha})\right\} = \eta'_\beta\left\{V_\alpha(\hat{\beta}) - \left(\frac{\partial\beta_\alpha}{\partial\alpha}\right)' I(\alpha)\frac{\partial\beta_\alpha}{\partial\alpha}\right\}\eta_\beta, \qquad (2.54)$$

where $\eta'_\beta = Cov_g\left\{\frac{\partial\ell_f(\alpha)}{\partial\alpha}, \ell_{gf}(\alpha_\beta, \beta)\right\}$.

Obtaining the expressions for this test statistic and its variance is a more time-consuming task than the corresponding procedures for the tests presented in the previous section. Moreover, the application of this test is not feasible in some cases. For example, when the null hypothesis H_f: $f(y, \alpha)$ is lognormal and the alternative H_g: $g(y, \beta)$ is exponential or Weibull in form, $S_f(\alpha) = 0$ (Rojas et al. 2008. See also Cox 1961, p. 119).

Example 2.12 (Rojas 2001) Consider the hypothesis testing of H_f (lognormal) against H_g (gamma) as in (2.12). We have

$$E_{\hat{\alpha}}\left[\ell_f(\hat{\alpha})\right] = -n\left[\ln(2\pi\psi'(\hat{\gamma}_2)) - \psi(\hat{\gamma}_2) - \ln\left(\frac{\hat{\gamma}_1}{\hat{\gamma}_2}\right) - \frac{1}{2}\right],$$
$$E_{\hat{\alpha}}\left[\ell_g(\hat{\gamma})\right] = -n\Big[\ln\left(\frac{\hat{\gamma}_1}{\hat{\gamma}_2}\right) + \ln\Gamma(\hat{\gamma}_2) - (\hat{\gamma}_2 - 1)\left\{\psi'(\hat{\gamma}_2) - \ln\left(\frac{\hat{\gamma}_1}{\hat{\gamma}_2}\right)\right\},$$
$$+\hat{\gamma}_2\Big]$$
$$E_{\hat{\alpha}}\left[\ell_f(\hat{\alpha}) - \ell_g(\hat{\gamma})\right] = -n\left[\ln(2\pi\psi'(\hat{\gamma}_2)) + \frac{1}{2} + \hat{\gamma}_2\psi(\hat{\gamma}_2) - \hat{\gamma}_2 - \ln\Gamma(\hat{\gamma}_2)\right], \qquad (2.55)$$

and thus,

$$S_f(\hat{\alpha}) = \frac{n}{2}\ln\left\{\frac{\psi'(\gamma_{2\hat{\alpha}})}{\psi'(\gamma_2)}\right\} + n\left[\gamma_{2\hat{\alpha}}\psi(\gamma_{2\hat{\alpha}}) - \hat{\gamma}_2\psi(\hat{\gamma}_2)\right]$$
$$= +n\ln\left\{\frac{\Gamma(\hat{\gamma}_2)}{\Gamma(\gamma_{2\hat{\alpha}})}\right\} + n\left(\hat{\gamma}_2 - \gamma_{2\hat{\alpha}}\right). \qquad (2.56)$$

When H_g is the null hypothesis and H_g is the alternative, we obtain

$$S_g(\hat{\beta}) = n\left[\ln\left\{\frac{\Gamma(\gamma_{2\alpha\hat{\gamma}})}{\Gamma(\gamma_{2\hat{\alpha}})}\right\} + \gamma_{2\alpha\hat{\gamma}}\left(\ln\gamma_{2\alpha\hat{\gamma}} + \frac{\alpha_{2\hat{\gamma}}}{2}\right)\right]$$
$$= -\gamma_{2\hat{\alpha}}\left(\ln\gamma_{2\hat{\alpha}} + \frac{\hat{\alpha}_2}{2}\right) - (\gamma_{2\alpha\hat{\gamma}} - 1)\alpha_{1\hat{\gamma}} + (\gamma_{2\hat{\alpha}} - 1)\hat{\alpha}_1 \qquad (2.57)$$
$$= +\gamma_{2\alpha\hat{\gamma}} - \gamma_{2\hat{\alpha}} + \ln\left\{\frac{\alpha_{2\hat{\gamma}}}{\hat{\alpha}_2}\right\} + \gamma_{1\hat{\gamma}} - \hat{\alpha}_1.$$

2.4.3 Test of the Nearest Alternative

Considering the measure of closeness (1.2), the Kullback–Leibler divergence or information criterion (KLIC) is

$$I(f, g) = \int \left\{\ell_f(\alpha) - \ell_g(\beta)\right\} f(y, \alpha)\mathrm{d}y.$$

Under Assumption 2.3,

$$E_\alpha \left[\frac{\partial}{\partial\beta}\ell_g(\beta_\alpha)\right] = 0.$$

Thus, it follows that β_α minimizes the KLIC.

Let us use a sample $y = (y_1, \ldots, y_n)$, where the y_i are iid random variables, to test the null hypothesis $H_f : f(y, \alpha)$ against the alternative hypothesis $H_g : g(y, \beta)$, and let us assume that the parameter vector β has a higher dimension than the parameter α of the null hypothesis. Shen (1982) proposed the use of the usual likelihood ratio test of the hypothesis $H_0 : g(y, \beta_\alpha)$ against the alternative $H_1 : g(y, \beta)$, where $g(y, \beta_\alpha)$ is the nearest alternative in $g(y, \beta)$ that is close to $f(y, \beta)$.

Example 2.13 (Shen 1982) Consider the hypotheses

H_f: exponential (β) and H_g: lognormal (α_1, α_2).

From Example 2.1, we have

$$\alpha_{1\beta} = \ln\beta + \psi(1), \qquad \alpha_{2\beta} = \psi'(1).$$

Under the null hypothesis, the lognormal likelihood function is proportional to

$$-\frac{1}{2}\log\psi'(1) - \frac{\sum(\log y_i - log\beta - \psi(1))^2}{2\psi'(1)}$$

and is maximized by taking $\beta = \exp\left\{\frac{1}{n}\sum\ln y_i - \psi(1)\right\}$. The likelihood ratio is therefore

$$LR = \frac{\left\{\psi'(1)\right\}^{-\frac{n}{n}} \exp\left\{\frac{-\sum(\ln y_i - n^{-1}\sum\ln y_i)^2}{2\psi'(1)}\right\}}{(\hat{\alpha}_2)^{-\frac{n}{2}}\exp\left(-\frac{n}{2}\right)}. \tag{2.58}$$

2.4.4 Test Based on the Moment Generating Function

Epps et al. (1982) derived a test for separate families of distributions based on the empirical generating function $M(t) = n^{-1}\sum e^{ty_j}$. They considered $\mu(t) = E\left(e^{tY}\right)$ and supposed that $\mu(t)$ exists and is equal to $\mu_f(t, \alpha)$. Under $H_f : \mu_f(t, \alpha)$, we have

$$E_\alpha\left\{M(t)\right\} = E\left\{\frac{M(t)}{H_f}\right\} = \mu_f(t,\alpha),$$
$$V_\alpha\left\{M(t)\right\} = n^{-1}\left\{\mu_f(2t,\alpha) - \mu_f^2(t,\alpha)\right\}, \tag{2.59}$$

where $\left\{M(t) - \mu_f(t,\alpha)\right\}\sqrt{n}$ is asymptotically normal for any t such that $0 < V_\alpha\left\{M(t)\right\} < \infty$. The authors extended this result for the testing of separate families of distributions when α is estimated using the maximum likelihood approach and by choosing t so as to maximize the power of the test of the separate families against the specified alternative $H_g : g(y,\beta)$.

Under the regularity conditions for maximum likelihood estimation,

$$Z_f(t,\hat{\alpha}) = \sqrt{n}\frac{M(t) - \mu_f(t,\hat{\alpha})}{\sigma_f(t,\hat{\alpha})} \tag{2.60}$$

converges in distribution to $N(0,1)$ for any t such that $0 < \sigma_f^2(t,\alpha) < \infty$, where

$$\sigma_f^2(t,\alpha) = \left\{\mu_f(2t,\alpha) - \mu_f^2(t,\alpha) - \left(\frac{\partial\mu_f(t,\alpha)}{\partial\alpha}\right)' I(\alpha)\left(\frac{\partial\mu_f(t,\alpha)}{\partial\alpha}\right)\right\}. \tag{2.61}$$

To consider the choice of the critical region and the power of the test, let us assume that $\hat{\alpha} \to \alpha_\beta$ under $H_g : \mu_g(t,\beta)$. Then, under H_g, $M(t) - \mu_f(t,\hat{\alpha})$ converges in probability to $\mu_g(t,\beta) - \mu_f(t.\alpha_\beta)$, and its asymptotic variance is

$$\begin{aligned}\sqrt{n}\sigma_g^2(t,\beta) = \sqrt{n}\Big\{&\mu_g(2t,\beta) - \mu_g^2(t,\beta)\\ &-2\left(\tfrac{\partial\mu_f(t,\alpha_\beta)}{\partial\alpha}\right)' I(\beta)\left(E_\beta\left\{e^{tY}\tfrac{\partial\ln f(y,\alpha_\beta)}{\partial\alpha}\right\}\right)\\ &-\left(I(\beta)\tfrac{\partial\mu_f(t,\alpha_\beta)}{\partial\alpha}\right)' I(\beta)E_\beta\left\{\left(\tfrac{\partial\ln f(y,\alpha_\beta)}{\partial\alpha}\right)'\left(\tfrac{\partial\ln f(y,\alpha_\beta)}{\partial\alpha}\right)\right\}\Big\}.\end{aligned} \tag{2.62}$$

This result is obtained from the results given in (2.8) Cox (1961).

Therefore, under H_g, the test statistic $Z_f(t,\hat{\alpha})$ in (2.60) is asymptotically distributed as $N\left\{k_1(t), k_2(t)\right\}$, where

$$k_1(t) = \sqrt{n}\frac{\mu_g(t,\beta) - \mu_f(t,\alpha_\beta)}{\sigma_f(t,\alpha_\beta)}, \tag{2.63a}$$

$$k_2(t) = \frac{\sigma_g(t,\beta)}{\sigma_f^2(t,\alpha_\beta)}. \tag{2.63b}$$

For a fixed large n, the power is maximized by choosing t to minimize

$$\pi(\alpha_{\hat{\beta}}, t) = k_2(t)^{-\frac{1}{2}}\left\{z_p - |\beta_1(t)|\right\}, \tag{2.64}$$

where z_p is the ordinate of the normal variate.

Example 2.14 (Epps et al. 1982) To test an exponential model versus a lognormal model, we can equivalently test the hypothesis that $X = \log Y$ is log-exponential (H_f) against the hypothesis of a normal distribution.

In this case, testing $H_f : \mu_f(t, \beta) = \beta^t \Gamma(t+1)$ against $H_g : \mu_g(t, \alpha_1, \alpha_2) = \exp\left(\alpha_1 t + \frac{1}{2}\alpha_2 t^2\right)$ implies that the test statistic is

$$Z_f(t, \beta) = \sqrt{n}\frac{M_X(t) - (\bar{Y})^t \Gamma(t+1)}{(\bar{Y})^t \left\{\Gamma(2t+1) + (1+t^2)\Gamma^2(t+1)\right\}} \; (-1 < t = 0, 1),$$

and under H_g, $\beta_\alpha = \exp\left(\alpha_1 + \frac{1}{2}\alpha_2\right)$.

$$\begin{aligned} \sigma_f^2(t, \beta_\alpha) &= \beta_\alpha^{2t}\left\{\Gamma(2t+1) - (1+t^2)\Gamma^2(t+1)\right\} \\ \sigma_g^2(t, \alpha) &= \beta_\alpha^{2t} e^{t(t-1)\alpha_2}\left\{\exp(t^2\alpha_2 - 1)\right\} \\ &= -\beta_\alpha^{2t} 2t\Gamma(t+1)\left\{e^{\frac{1}{2}(t+1)t\alpha_2} - e^{\frac{1}{2}t(t-1)\alpha_2}\right\} \\ &= \beta_\alpha^{2t} t^2 \Gamma^2(t+1)(e^{\alpha_2} - 1). \end{aligned} \tag{2.65}$$

For further details and results, see Epps et al. (1982).

2.4.5 Two Further Tests

Here, we briefly discuss several other tests for separate families of hypotheses.

First, the Vuong (1989) procedure for discriminating separate hypotheses H_f and H_g considers the null hypothesis

$$\begin{aligned} H_0 &: E_\alpha\left[\ell_f(\alpha) - \ell_g(\beta)\right] = 0 \text{ (both models are equivalent)} \\ &\quad \text{against} \\ H_f &: E_\alpha\left[\ell_f(\alpha) - \ell_g(\beta)\right] > 0 \; (H_f \text{ is superior to } H_g) \\ &\quad \text{or} \\ H_g &: E_\alpha\left[\ell_f(\alpha) - \ell_g(\beta)\right] < 0 \; (H_g \text{ is superior to } H_f). \end{aligned} \tag{2.66}$$

The test statistic proposed by Vuong is

- an unadjusted likelihood ratio statistic

$$\sqrt{n}\frac{\ell_f(\hat{\alpha}) - \ell_g(\hat{\beta})}{\hat{v}_n}, \tag{2.67}$$

where

$$\hat{v}_n = \left[\frac{1}{n}\sum_{i=1}^{n}\left\{\ln\frac{f(y_i, \hat{\alpha})}{g(y_i, \hat{\beta})}\right\}^2 - \left\{\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(y_i, \hat{\alpha})}{g(y_i, \hat{\beta})}\right\}^2\right]^{\frac{1}{2}},$$

or

- an adjusted likelihood ratio statistic

$$\sqrt{n}\left[\left\{\ell_f(\hat{\alpha}) - \ell_g(\hat{\beta})\right\} - \xi(f, g)\right] \tag{2.68}$$

with a correction in the denominator $\hat{v}_n$, where $\xi(f, g)$ is a correction factor that depends on the characteristics of the models, such as their numbers of parameters. Examples of correction factors include $\xi(f, g) = p - q$ and $\frac{p-q}{2} \ln n$, where p and q are the numbers of parameters of $f(y, \alpha)$ and $g(y, \beta)$, respectively.

Vuong's hypotheses generalize the discussion of Hottelling (1940) regarding the hypothesis that two alternative predictors in linear regression are equally effective (Cox 2013). In the context of separate models, geometric interpretation and further generalizations of Hottelling's prediction problem are discussed in Efron (1984).

Smith (1992) proposed another test statistic for nonnested regression models estimated using the generalized method of moment (GMM). For $Y = (y_t, y_{t-1}, \ldots, y_1, y_0)$, where y_0 represents the initial condition of the process, he considered two hypotheses:

$$\begin{aligned} H_f &: \; E_f\left[f(y_t, \alpha)\right] = 0, \\ H_g &: \; E_g\left[g(y_t, \beta)\right] = 0, \end{aligned} \tag{2.69}$$

where $f(y_t, \alpha)$ and $g(y_t, \beta)$ are k_f and k_g continuous differentiable vector functions of the p_f and p_g vectors of parameters α and β, respectively, such that $k_f > p_f$ and $k_g > p_g$ (here, f and g are not densities).

To test the null hypothesis H_f against the alternative hypothesis H_g, Smith proposed the test statistic below based on some results of GMM estimation (Kent 1986):

$$\hat{\Gamma} - \tilde{\Gamma}, \tag{2.70}$$

where $\hat{\Gamma}$ is a function of $f(y_t, \hat{\alpha})$ and $g(y_t, \hat{\beta})$ and $\tilde{\Gamma}$ is the probability limit of $\hat{\Gamma}$ under the hypothesis H_f. Here, $\hat{\alpha}$ and $\hat{\beta}$ are the GMM estimators of α and β, respectively, and β_α is the probability limit of the GMM estimator $\hat{\beta}$ under H_f. The procedure is a GMM analog of Cox's MLE procedure.

Simulation results comparing the performance of this test with the Cox test have been presented by Arkonac and Higgins (1995).

Using the general concept of (2.69), Otsu et al. (2012) proposed a test employing the generalized empirical likelihood (GEL), which includes the GMM as a special case. Monte Carlo experiments have also been presented for the test of the logistic model as the null model against the Gumbel and Burr models.

2.5 Efficiencies of False Separate Models

2.5.1 Introduction

The consequences of using an incorrect model are investigated in this section. A recent discussion of Cox's original paper illustrates the importance of this topic. Cox (2013) stated,

> Mathematically the most fruitful part of the paper is a side issue: the study of the distribution of a maximum likelihood estimate when the model fitted and the data-generating model are not the same. What is now called the sandwich formula arises in a number of quite different contexts.

The participants in that discussion emphasized the relation between these ideas and later developments such as robustness, misspecification and encompassing.

The results of Kent (1982) concerning the use of a false model in the Holy Trinity of tests—the likelihood ratio (Wilks 1938), Wald (1943) and Rao (1947) tests—are very important. Kent's results can also possibly be extended to the asymptotically equivalent test of Terrel (2002). It is remarkable that Terrel's simple equivalent test was developed only recently.

2.5.2 Efficiency of a false regression model

In this discussion, we are interested only in the regression coefficients and the properties of their estimators. For the models treated in Example 2.3 and the corresponding probability limits, the estimators of the m regression coefficients are asymptotically consistent, independent of distributional assumptions. Therefore, the asymptotic variances are of primary interest for the comparison of the estimators obtained from alternative models.

Suppose that the true hypothesis is H_f, that the model specified by H_g is used, and that α^* and β^* are the components of α and β, respectively, that correspond to $m > 1$ regression coefficients.

The efficiency of a false model is measured in terms of the ratio of determinants,

$$\text{eff}_\alpha(\hat{\beta}^*) = \frac{|V_\alpha(\hat{\alpha}^*)|^{1/m}}{|V_\alpha(\hat{\beta}^*)|^{1/m}} \quad (m \geq 1), \tag{2.71}$$

and provides insight into the results obtained using that false model.

It is also useful to find the element that corresponds to

$$V^*_\alpha\left(\beta^*\right) = n^{-1}\text{p}\lim_f \left[n\left\{E_\beta(G_{\beta^{*\prime}\beta^*})\right\}^{-1}\right]_{\beta=\hat{\beta}}, \tag{2.72}$$

the probability limit under H_f of the false estimator for the covariance matrix of $\hat{\beta}^*$, which is used when it is not known that the model is wrong.

Finally, we note a general simplification of our models that is brought about by the parameterization (2.19) of the z_i. With the notation of Example 2.3, it can easily be shown (Cox and Hinkley 1968) that for log-linear models, the matrices

$$E_\alpha\left(\frac{\partial \ln f(y,\alpha)}{\partial\alpha}\right),\ E_\alpha\left(\frac{\partial^2 \ln g(y,\beta_\alpha)}{\partial\beta'\partial\beta}\right) \text{ and } E_\alpha\left(\frac{\partial \ln g(y,\beta)}{\partial\beta}\frac{\partial \ln g(y,\beta)}{\partial\beta}\right)$$

all take the general form

$$\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}.$$

The submatrices A are square matrices of the expected values of the derivatives corresponding to the general mean and the shape or scale of log y_i. The submatrices B are matrices corresponding to the regression coefficients, which can obtained using the results outlined in Appendix A and in Example 2.3 Consequently, it is necessary only to determine these submatrices in order to evaluate (2.71) and (2.72).

(i) Lognormal regression model

Suppose that the true model is H_L. The asymptotic covariance matrix of $\hat{a}$ is then $V_L(\hat{a}) \sim (Z'Z)^{-1}\alpha_2$. The consequences of using other models are discussed below.

If the Weibull regression model is falsely assumed, then we have

$$E_L(\ell_{W,b'b}) = -Z'Z/\alpha_2,\quad E_L(\ell_{W,b}\ell_{W,b}) = Z'Z(e-1)/\alpha_2;$$

thus, (2.71) and (2.72) imply that

$$V_L^*(\hat{b}) \sim (Z'Z)^{-1}\alpha_2,\quad V_L(\hat{b}) \sim (Z'Z)^{-1}(e-1)\alpha_2,$$

and $\mathrm{eff}_L(\hat{b}) = (e-1)^{-1} = 0.58$. Thus, $V_L(\hat{b}_j)$ is 72 % higher than its stated estimate $V_L^*(\hat{b}_j)$.

If the gamma regression model is falsely assumed, then we have

$$E_L(\ell_{G,c'c}) = -Z'Z_{\gamma 2L},\quad E_L(\ell_{G,c}\ell_{G,c'}) = Z'Z\,(e^{\alpha_2}-1)\,\gamma_{2L}^2;$$

therefore, $V_L^*(\hat{c}) \sim (Z'Z)^{-1}\gamma_{2L}^{-1}$, $V_L(\hat{c}) \sim (Z'Z)^{-1}\,(e^{\alpha_2}-1)$, and $\mathrm{eff}_L(\hat{c}) = \alpha_2/\,(e^{\alpha_2}-1)$.
The efficiency rapidly decreases as α_2 increases. Thus, for $\alpha_2 = 0.2$, 1.0, and 2.0, the efficiencies are 0.9, 0.58, and 0.27, respectively. Furthermore, $\mathrm{eff}_L(\hat{c})$ approaches 1 as $\alpha_2 \to 0$. This is because as α_2 tends to zero, the lognormal distribution approaches a normal distribution. For a normal distribution with

mean $\exp(z\alpha)$, the maximum likelihood equation for α is the same as that for the gamma regression model.

(ii) Weibull regression model

Suppose that the true model is H_W. The asymptotic covariance matrix of $\hat{b}$ is $V_W(\hat{b}) \sim (Z'Z)^{-1}/\beta_2^2$.

If the lognormal regression model is falsely assumed, then we have

$$E_W(\ell_{L,a'a}) = -Z'Z\beta_2^2/\psi'(1), \quad E_W(\ell_{L,a}\ell_{L,a'}) = Z'Z\beta_2^2/\psi'(1);$$

therefore,

$$V_W^*(\hat{a}) \sim (Z'Z)^{-1}\psi'(1)/\beta_2^2, \quad V_W(\hat{a}) \sim (Z'Z)^{-1}\psi'(1)/\beta_2^2,$$

and $\mathrm{eff}_W(\hat{a}) = \psi'(1)^{-1} = 0.61$, where $\psi(x) = d\log\Gamma(x)/dx$. Here, $V_W^*(\hat{a})$ shows that a correct estimate of the variance of $\hat{a}_j$, the least squares estimator of b_j, is stated.

If the gamma regression model is falsely assumed, then we have $E_W(\ell_{G,c'c})' = -Z'Z\gamma_{2W}$ and $E_W(\ell_{G,c}\ell_{G,c'}) = Z'Z\gamma_{2W}^2\eta^2$; therefore, $V_W^*(\hat{c}) \sim (Z'Z)^{-1}/\gamma_{2W}$, $V_W(\hat{c}) \sim (Z'Z)^{-1}\eta^2$, and

$$\mathrm{eff}_W(\hat{c}) = (\beta_2\eta)^{-2},$$

where $\eta^2 = \Gamma(2\beta_2^{-1}+1)/\Gamma^2(\beta_2^{-1}+1) - 1$ is the square of the coefficient of variation of a Weibull distribution with shape parameter β_2. Table 2.6 lists the efficiency and other values of interest. The efficiency is high for β_2 near 1, as expected, and it decreases for β_2 far from 1. These results for γ_{2W}, η^2 and $V_W^*(\hat{c})$ suggest that an underestimate or an overestimate of $V_W(\hat{c})$ is obtained when $\beta_2 < 1$ or when $\beta_2 > 1$, respectively.

(iii) Gamma regression model

Suppose that the true model is H_G. Then, the asymptotic covariance matrix of $\hat{c}$ is $V_G(\hat{c}) \sim (Z'Z)^{-1}/\gamma_2$.

If the lognormal regression model is falsely assumed, then we have

$$E_G(\ell_{L,a'a}) = -Z'Z/\psi'(\gamma_2), \quad E_G(\ell_{L,a}\ell_{L,a'}) = Z'Z/\psi'(\gamma_2);$$

therefore, $V_G^*(\hat{a}) \sim (Z'Z)^{-1}\psi'(\gamma_2)$, $V_G(\hat{a}) \sim (Z'Z)^{-1}\psi'(\gamma_2)$, and

$$\mathrm{eff}_G(\hat{a}) = \left\{\gamma_2\psi'(\gamma_2)\right\}^{-1}.$$

Table 2.6 Efficiency of the gamma regression model when H_w is true

β_2	0.4	0.6	0.8	1.2	2.0	5.0
γ_{2W}	0.266	0.468	0.712	1.333	3.131	16.612
η^2	9.865	3.091	1.589	0.699	0.273	0.052
eff	0.63	0.90	0.98	0.99	0.92	0.76

The efficiency approaches 1 as γ_2 increases. This is because as γ_2 increases, the gamma distribution approaches a lognormal distribution. When γ_2 decreases to zero, the efficiency also tends to zero. For further values, see Cox and Hinkley (1968). In this situation, $V_G^*(\hat{a})$ shows that a correct estimate of the variance of $\hat{a}_j$, the least squares estimator of c_j, is stated.

If the Weibull regression model is falsely assumed, then we have

$$E_G(\ell_{W,b'b}) = -Z'Z\beta_{2G}^2, \quad E_G(\ell_{W,b}\ell_{W,b'}) = Z'Z\beta_{2G}^2\eta^2;$$

therefore, $V_G^* \sim (Z'Z)^{-1}/\beta_{2G}^2$, $V_G(\hat{c}) \sim (Z'Z)^{-1}(\eta/\beta_{2G})^2$, and

$$\text{eff}_W(\hat{b}) = (\beta_{2G}/\eta)^2/\gamma_2,$$

where $\eta^2 = \Gamma(2\beta_{2G} + \gamma_2)\Gamma(\gamma_2)/\Gamma^2(\beta_{2G} + \gamma_2) - 1$
is the square of the coefficient of variation of $V^{\beta_{2G}}$, where V is a gamma distribution with shape parameter γ_2.

Table 2.7 presents the efficiency and other values of interest. As expected, the efficiency is high for γ_2 near 1 and decreases for γ_2 far from 1. These results for η^2 and $V_G^*(\hat{b})$ suggest that $V_G(\hat{b})$ is overestimated if $\gamma_2 < 1$ and underestimated if $\gamma_2 > 1$.

(iv) Special case: Exponential regression model
The results for the exponential regression model can be inferred from the previous results. It is easy to see that the maximum likelihood estimators for the parameters of the exponential regression model are the same as those for the gamma regression model given γ_2. Therefore, when the exponential regression model is a false model, the results are the same as those for the gamma regression model, omitting the shape factor γ_2. When the exponential regression model is the true model, the results can be obtained from those presented for the gamma and Weibull models with $\beta_2 = \gamma_2 = 1$.

Table 2.7 Efficiency of the Weibull regression when H_G is true

β_2	0.4	0.6	0.8	1.2	2.0	5.0
γ_{2G}	0.534	0.718	0.870	1.115	1.482	2.370
η^2	0.807	0.892	0.951	1.039	1.142	1.304
eff	0.89	0.96	0.99	0.997	0.96	0.86

2.6 Properties and Comparisons

2.6.1 Asymptotic Power

When studying the asymptotic power function of consistent tests, the type I error is held fixed but the alternative hypothesis is allowed to approach the null hypothesis. When the hypotheses are nested (say H_f is nested within H_g), there is no difficulty. Pesaran (1984) also used this approach for the case in which the hypotheses H_f and H_g are partially nonnested.

In the case of nonnested hypotheses, however, this approach is not possible by definition, and an alternative method proposed by Pesaran (1984) is to use the Bahadur approach, in which the alternative hypothesis fixed is held fixed but the type I error is allowed to tend to zero as the sample size increases. The significance levels of the tests for a fixed power are compared against a specific alternative.

Let $\hat{\alpha}_n$ denote the asymptotic significance level of a test. Bahadur calls the quantity

$$\lim_{n\to\infty}\left(-\frac{2}{n}\ln\hat{\alpha}_n\right) \tag{2.73}$$

the asymptotic or "approximate slope" of the test, and a test is considered asymptotically efficient relative to another if its approximate slope is greater.

Pesaran (1984) extended Bahadur's result to the case of separate hypotheses. He established that if a test statistic Z^2 asymptotically possesses a central $\chi^2_{(n)}$ distribution under the null hypothesis H_f, then

$$\lim_{n\to\infty}\left(-\frac{2}{n}\ln\hat{\alpha}_n\right) = p\lim_{n\to\infty}\left\{n^{-1}Z^2|H_g\right\}. \tag{2.74}$$

Example 2.15 (Pesaran 1984) Consider a test of the null hypothesis H_E against the alternative H_L. Using expression (2.9) from Example 2.1 and item (i) from Example 2.7, Cox's and Atkinson's test statistics under H_L are, respectively,

$$\begin{aligned} Z_{LE}(C) &= \frac{\sqrt{n}\left(\ln\hat{\beta}-\hat{\alpha}_1-\hat{\alpha}_2/2\right)}{\sqrt{e^{\hat{\alpha}_2}-1-\hat{\alpha}_2-\frac{1}{2}\hat{\alpha}_2^2}}, \\ Z_{LE}(A) &= \frac{\sqrt{n}\left[\hat{\alpha}_1\exp(-\hat{\alpha}_1-\hat{\alpha}_2)-1\right]}{\sqrt{e^{\hat{\alpha}_2}-1-\hat{\alpha}_2-\frac{1}{2}\hat{\alpha}_2^2}}. \end{aligned} \tag{2.75}$$

Meanwhile, under the alternative H_E,
$\alpha_1 \xrightarrow{p} \psi(1)+\ln\beta, \alpha_2 \xrightarrow{p} \psi'(1)$ and $\hat{\beta} \rightarrow \beta$ ($\psi(1) = -0.5772$, $\psi'(1) = 1.6449$).

Upon substituting these results into expression (2.74), the slope of the Cox test is 0.0509 and that of the Atkinson test is 0.040, which implies that the Cox test is 27 % (0.0509/0.040) more asymptotically efficient than the Atkinson test.

Suppose that the roles of H_L and H_E are reversed, that is, H_E is the null hypothesis and H_L is the alternative hypothesis; then, using Eq. (2.10) from Example 2.1 and item (ii) from Example 2.6, Cox's and Atkinson's test statistics are, respectively,

$$\begin{aligned} Z_{EL}(C) &= \left(\tfrac{n}{0.2834}\right)^{\frac{1}{2}} \left\{\hat{\alpha}_1 - \psi(1) + \ln\hat{\beta} + \tfrac{1}{2}\ln\hat{\alpha}_2 - \tfrac{1}{2}\ln\psi'(1)\right\}, \\ Z_{EL}(A) &= \left(\tfrac{n}{0.2834}\right)^{\frac{1}{2}} \left\{\hat{\alpha}_1 - \psi(1) - \ln\hat{\beta}\right. \\ &\quad \left. +\tfrac{1}{2}\psi'(1)\left[\hat{\alpha}_2 - \psi'(1) + (\hat{\alpha}_1 - \psi(1) - \ln\hat{\beta})^2\right]\right\}. \end{aligned} \tag{2.76}$$

Under H_L, $\hat{\beta} \xrightarrow{p} \exp\left\{\alpha_1 + \frac{\alpha_2}{2}\right\}$, $\hat{\alpha}_1 \xrightarrow{p} \alpha_1$, and $\hat{\alpha} \rightarrow \alpha_2$.

Substituting these results into (2.74), we obtain the following as $n \rightarrow \infty$:

$$\begin{aligned} \text{p}\lim\left(n^{-1}Z_{EL}^2(C|H_L)\right) &= 0.882(\alpha_2 - \ln\alpha_2 - 0.6567)^2, \\ \text{p}\lim\left(n^{-1}Z_{EL}^2(A|H_L)\right) &= (0.1427\alpha_2^2 - 0.6978\alpha_2 + 0.3352)^2. \end{aligned} \tag{2.77}$$

The Bahadur asymptotic efficiency of the two tests varies with the parameter α_2. The Cox test is always more efficient than the Atkinson test, because the Atkinson test is only consistent for values of α_2 inside the interval (0.5401, 4.3484), as shown in Pereira (1977b). Note that (Pesaran 1984) statement that the Atkinson test is inconsistent only for values of $\beta = 0.54$ and $\beta = 4.35$ is incorrect.

Now, we will obtain the power results for Shen's test presented in Sect. 2.4.4. To calculate the approximate slope of the test, namely, the limit of (2.58) under H_g: lognormal, we note that as $\hat{\alpha}_1 \rightarrow \alpha_1$ and $\hat{\alpha}_2 \rightarrow \alpha_2$ from (2.74), we obtain the expression

$$\lim_{n\rightarrow\infty} \frac{1}{n}\left[\frac{\alpha_2}{\psi'(1)} \exp\left\{\frac{\psi'(1) - \alpha_2}{\psi'(1)}\right\}\right]^{\frac{n}{2}} = 0. \tag{2.78}$$

For $\alpha_2 \geq 0$, the expression inside the brackets is always less than 1. Therefore, the Cox test is also more efficient in this case.

2.6.2 *Monte Carlo Comparison and Behavior*

Empirical results regarding a comparison of the Cox and Atkinson tests and the adequacy of asymptotic results for finite samples are discussed in this section. A general pattern observed in the simulation results is described. Only simulations of the lognormal and Weibull distributions are presented because the maximum likelihood ratio is independent of the parameters in these cases.

Table 2.8 Null distributions of $T_{LW}(C)$ and $T_{LW}(A)$

n	$T_{LW}(\cdot)$	$\mu_1\{T_{LW}(\cdot)\|H_L\}$	$\mu_2\{T_{LW}(\cdot)\|H_L\}$	$\gamma_1\{T_{LW}(\cdot)\|H_L\}$	$\beta_2\{T_{LW}(\cdot)\|H_L\}$
20	C	−0.261	0.502	0.090	3.387
	A	−0.118	0.503	1.665	8.366
50	C	−0.232	0.686	0.167	3.131
	A	−0.103	0.723	1.433	8.033
100	C	−0.198	0.758	0.329	3.197
	A	−0.092	0.818	1.186	5.602
150	C	−0.163	0.789	0.298	2.867
	A	−0.072	0.832	0.880	4.000
200	C	−0.142	0.805	0.355	3.368
	A	−0.058	0.882	1.088	5.511

Results from 1000 trials

Table 2.9 Distributions of $T_{LW}(C)$ and $T_{LW}(A)$ under the alternative H_W

n	$T_{LW}(\cdot)$	$\mu_1\{T_{LW}(\cdot)\|H_W\}$	$\mu_2\{T_{LW}(\cdot)\|H_W\}$	$\gamma_1\{T_{LW}(\cdot)\|H_W\}$	$\beta_2\{T_{LW}(\cdot)\|H_W\}$
20	C	−1.387	0.720	−0.492	3.459
	A	−0.913	0.215	0.510	3.776
50	C	−2.419	1.003	−0.562	3.950
	A	−1.638	0.266	0.155	3.519
100	C	−3.584	1.148	−0.371	3.406
	A	−2.445	0.297	0.126	3.502
150	C	−4.436	1.256	−0.283	3.391
	A	−3.038	0.324	−0.116	3.415
200	C	−5.119	1.257	0.395	3.344
	A	−3.522	0.323	0.099	3.162

Results from 1000 trials

The results of Pereira (1981a) are presented in Tables 2.8, 2.9, 2.10, 2.11, 2.12, 2.13, 2.14 and 2.15. It can be seen from these results that the Atkinson test statistic approaches its asymptotic mean and variance faster than does the Cox test statistic, whereas the reverse occurs for the third and fourth moments.

The same pattern was observed by (Pereira 1976, 1977a, 1981a) for tests involving pairwise comparisons of the lognormal, gamma, Weibull and exponential distributions. In fact, this is a general result, which will be discussed in the next section.

Our simulation results agree with those of Jackson (1968) and Jackson (1969).

Some pioneering simulations related to these previous works are those of Dumonceaux and Antle (1973), and Dumonceaux et al. (1973a).

Rojas (2001) presented several short simulations to check the approach to the asymptotic distribution of the test presented in Sect. 2.4.3 and also to check the probability of correct selection using Lindsey's procedure presented in Sect. 4.2.

Additional simulation results can be found in Pereira (2005, 2010) and the references therein.

Table 2.10 Null distributions of $T_{WL}(C)$ and $T_{WL}(A)$

n	$T_{WL}(\cdot)$	$\mu_1\{T_{WL}(\cdot)\|H_W\}$	$\mu_2\{T_{WL}(\cdot)\|H_W\}$	$\gamma_1\{T_{WL}(\cdot)\|H_W\}$	$\beta_2\{T_{WL}(\cdot)\|H_W\}$
20	C	−0.224	0.555	0.492	3.459
	A	−0.084	0.665	1.777	7.723
50	C	−0.094	0.918	0.512	3.480
	A	−0.043	0.089	1.406	6.059
100	C	−0.078	0.884	0.371	3.406
	A	0.011	0.957	0.984	4.481
150	C	−0.055	0.967	0.283	3.391
	A	0.023	1.018	0.824	4.335
200	C	−0.067	0.968	0.395	3.344
	A	−0.001	1.016	0.815	4.111

Results from 1000 trials

Table 2.11 Distributions of $T_{WL}(C)$ and $T_{WL}(A)$ under the alternative H_L

n	$T_{WL}(\cdot)$	$\mu_1\{T_{WL}(\cdot)\|H_L\}$	$\mu_2\{T_{WL}(\cdot)\|H_L\}$	$\gamma_1\{T_{WL}(\cdot)\|H_L\}$	$\beta_2\{T_{WL}(\cdot)\|H_L\}$
20	C	−1.213	0.387	−0.090	3.387
	A	−0.858	0.122	1.380	6.072
50	C	−2.076	0.528	−0.167	3.131
	A	−1.451	0.118	0.857	3.625
100	C	−3.050	0.584	−0.329	3.197
	A	−2.120	0.104	0.581	3.379
150	C	−3.806	0.608	−0.298	2.867
	A	−2.631	0.098	0.407	3.027
200	C	−4.433	0.670	0.546	4.164
	A	−3.049	0.097	0.470	3.137

Results from 1000 trials

Table 2.12 Null: lognormal; alternative: Weibull. Tests: $T_{LW}(C)$ and $T_{LW}(A)$. Power at $t = -1.64$ and $t = -1.28$

n	$T_{LW}(\cdot)$	Power function	
		$SL = 0.05$	$SL = 0.10$
20	C	0.344	0.506
	A	0.045	0.217
50	C	0.771	0.887
	A	0.511	0.756
100	C	0.974	0.986
	A	0.940	0.977
150	C	0.994	0.997
	A	0.989	0.996
200	C	1.000	1.000
	A	1.000	1.000

Results from 1000 trials

Table 2.13 Null: lognormal; alternative: Weibull. Tests: $T_{LW}(C)$ and $T_{LW}(A)$. One-sided significance levels at $t = -1.64$ and $t = -1.28$

n	$T_{LW}(\cdot)$	Significance level	
		$SL = 0.05$	$SL = 0.10$
20	C	0.022	0.071
	A	0.000	0.010
50	C	0.043	0.106
	A	0.001	0.042
100	C	0.040	0.093
	A	0.008	0.051
150	C	0.032	0.096
	A	0.009	0.053
200	C	0.041	0.101
	A	0.016	0.067

Results from 1000 trials

Table 2.14 Null: Weibull; alternative: lognormal. Tests: $T_{WL}(C)$ and $T_{WL}(A)$. Power at $t = -1.64$ and $t = -1.28$

n	$T_{WL}(\cdot)$	Power function	
		$SL = 0.05$	$SL = 0.10$
20	C	0.231	0.447
	A	0.000	0.057
50	C	0.738	0.860
	A	0.330	0.751
100	C	0.973	0.996
	A	0.925	0.986
150	C	0.999	1.000
	A	0.996	1.000
200	C	1.000	1.000
	A	1.000	1.000

Results from 1000 trials

2.6.3 Test Consistency and Finite-Sample Results

A test of a hypothesis H_f against a class of alternatives H_g is said to be consistent if, when any member of H_g holds, the probability of rejecting H_f tends to one as the sample size tends to infinity.

As mentioned in Sect. 2.6.1, the Atkinson test statistic T_{fg} is not consistent when H_f is the exponential distribution and is tested against H_g, the lognormal distribution. Pereira (1977a) has shown that whereas the Cox test is always consistent, the Atkinson test may be inconsistent and therefore should be used only after verifying its consistency under the alternative hypothesis of interest. Fisher and McAleer

Table 2.15 Null: Weibull; alternative: lognormal. Tests: $T_{WL}(C)$ and $T_{WL}(A)$. One-sided significance levels at $t = -1.64$ and $t = -1.28$

n	$T_{WL}(\cdot)$	Significance Level	
		$SL = 0.05$	$SL = 0.10$
20	C	0.016	0.062
	A	0.000	0.000
50	C	0.023	0.078
	A	0.003	0.025
100	C	0.034	0.084
	A	0.015	0.047
150	C	0.045	0.087
	A	0.020	0.060
200	C	0.043	0.103
	A	0.020	0.076

Results from 1000 trials

(1981) have shown that for the testing of alternative regression models, the Atkinson test is consistent.

The small sample studies mentioned in Sect. 2.6.2 indicate a regular pattern in the comparison of the Cox and Atkinson tests.

To consider the behavior of the lower moments, the test statistics can be written as follows:

$$\begin{aligned} T_f(C) &= \ell_f(\hat{\alpha}) - \ell_g(\hat{\beta}) - E_{\hat{\alpha}}\left\{\ell_f(\hat{\alpha}) - \ell_g(\beta_{\hat{\alpha}})\right\}, \\ T_f(A) &= \ell_f(\hat{\alpha}) - \ell_g(\beta_{\hat{\alpha}}) - E_{\hat{\alpha}}\left\{\ell_f(\hat{\alpha}) - \ell_g(\beta_{\hat{\alpha}})\right\}. \end{aligned} \tag{2.79}$$

As noted by Atkinson (1970, p. 335), when α is estimated, both statistics will be biased, but $T_{fg}(A)$ will be less biased. It thus follows that in the Cox and Atkinson tests, it is expected that the asymptotic variance will be approached more rapidly for $T_{fg}(A)$ than for $T_{fg}(C)$ because in theory, the variance is calculated as if both statistics are unbiased.

Let us now consider the approach to normality of the distributions of $T_{fg}(C)$ and $T_{fg}(A)$. This behavior is related to the third- and fourth-order central moments. The test statistics can also be written as

$$\begin{aligned} T_{fg}(C) &= \ell_f(\hat{\alpha}) - \ell_g(\hat{\beta}) - E_{\hat{\alpha}}\left\{\ell_f(\alpha) - \ell_g(\beta_{\alpha})\right\}, \\ T_{fg}(A) &= \ell_f(\hat{\alpha}) - \ell_g(\beta_{\hat{\alpha}}) - E_{\hat{\alpha}}\left\{\ell_f(\alpha) - \ell_g(\beta_{\alpha})\right\}, \end{aligned} \tag{2.80}$$

where $\ell_f(\alpha) = \log L_f(\alpha, y)$, $\ell_g(\beta) = \log L_g(\beta, y)$ and E_α denotes the expectation value under H_f. The statistics given in (2.80) can be approximated by expanding $E_{\hat{\alpha}}\left\{\ell_f(\alpha)\right\}$ and $E_{\hat{\alpha}}\left\{\ell_g(\beta)\right\}$ around α, $\ell_f(\alpha)$ around $\hat{\alpha}$ and $\ell_g(\beta_\alpha)$ around $\hat{\beta}$, and $\beta_{\hat{\alpha}}$ to obtain

$$\begin{aligned} T_{fg}(C) &= T_{fg} + U_n, \\ T_f(A) &= T_{fg} + U_n + (\beta_\alpha - \beta_{\hat{\alpha}})\tfrac{\partial \ell_g(\beta)}{\partial \beta}, \end{aligned} \tag{2.81}$$

where T_{fg} (Cox 1962, Eq. (2.80)) is the sum of the deviations of $\log f(y_i, \alpha) - \log g(y_i, \beta_\alpha)$ from its regression on $\partial \log f(y_i, \alpha)/\partial\alpha$ and is of order $\sqrt{n}$ in probability, whereas the other terms are of order one in probability.

T_{fg} is a sum of iid random variables of zero mean, and therefore, a generally strong central limit effect can be expected to apply, unless, of course, the individual components have a markedly badly behaved distribution. The properties of U_n depend on the particular application, but U_n will often approach its limiting form quite rapidly. In any case, it affects both $T_{fg}(A)$ and $T_{fg}(C)$. The last term of $T_{fg}(A)$ in (2.81), at least in some applications, may follow a markedly non-normal distribution in samples of moderate size, and it is the poor behavior of this term that accounts for the slower convergence of the distribution of $T_{fg}(A)$. In particular, for some of the distributions investigated by Pereira (1976, 1977a, 1978), $\partial\ell_g(\beta_{\hat{\alpha}})$ requires a large sample size to become relatively small.

Under the null hypothesis, the C statistics should be preferable in terms of skewness and kurtosis. Therefore, from a practical point of view, the C statistics are generally recommended because corrections for lower order moments are considerably more easily obtained.

Example 2.16 (Pereira 1976, 1978) For the test presented in Sect. 2.6.2 of the lognormal distribution against the Weibull distribution, the term

$$(\beta_\alpha - \beta_{\hat{\alpha}})\frac{\partial\ell_g(\beta)}{\partial\beta}, \tag{2.82}$$

which differentiates the Atkinson test from the Cox test, takes a different form for each test as follows:

(i) For $T_{LW}(A)$, one of the terms in expression (2.82) is

$$\frac{\partial}{\partial\beta_1}\ell_W(\beta_{1\hat{\alpha}}, \beta_{2\hat{\alpha}}) = \frac{\beta_{2\hat{\alpha}}}{\beta_{1\hat{\alpha}}}\sum_{i=1}^{n}\left\{\left(\frac{y_i}{\beta_{1\hat{\alpha}}^{\beta_{2\hat{\alpha}}}}\right) - 1\right\}. \tag{2.83}$$

From the properties of the lognormal distribution, $\frac{y_i^{\beta_{2\hat{\alpha}}}}{\beta_{1\hat{\alpha}}^{\beta_{2\hat{\alpha}}}}$ has a lognormal distribution with $\alpha_1 = -1/2$ and $\alpha_2 = 1$. Therefore, when α_2 is large, the sample mean is an inefficient estimator of the mean of the lognormal distribution because a large sample size is required to make (2.83) negligible.

(ii) For $T_{WL}(C)$, the terms of (2.82) become

$$\frac{\partial}{\partial\alpha_1}\ell_L(\alpha_{1\hat{\beta}}, \alpha_{2\hat{\beta}}) = \frac{1}{\alpha_{2\hat{\beta}}}\sum_{i=1}^{n}(\log y_i - \alpha_{1\hat{\beta}}), \tag{2.84}$$

$$\frac{\partial}{\partial\alpha_2}\ell_L(\alpha_{1\hat{\beta}}, \alpha_{2\hat{\beta}}) = -\frac{n}{2\alpha_{2\hat{\beta}}} + \frac{1}{2\alpha_{2\hat{\beta}}^2}\sum_{i=1}^{n}(\log y_i - \alpha_{1\hat{\beta}})^2. \tag{2.85}$$

It is known that for the extreme value distribution, the efficiency of the method of moments in relation to the maximum likelihood method in estimating the location parameter is approximately 95%, and for the scale parameter, this efficiency is approximately 55 %. Therefore, at least (2.85) will require a large sample size to become negligible.

2.7 Bibliographic Notes

The original work on the efficiency of incorrect models was performed by Cox and Hinkley (1968). That paper focused on the efficiency of least squares estimates in relation to the Pearson Type VII and gamma distributions. Gould and Lawless (1988) presented general results on the consistency and efficiency of regression coefficient estimates in location–scale models. Cox (2013) and discussants noted that these results are related and pioneered the recent work on misspecification and what is known as the "Sandwich" formula for covariance matrices.

Procedures for censored data have been addressed by Slud (1983), Fine (2002) and Dey and Kundu (2012a). Kundu and associates have also applied Cox's results to binary comparisons, multiple tests and bivariate distributions; see Gupta and Kundu (2004), Kundu (2005), Kundu and Raqab (2007), Dey and Kund (2009, 2012a, 2012b) and references in their previous works.

The application of Cox's results for testing normality versus lognormality was studied by Kotz (1973). Recent works and references to applications involving the testing of linear versus log-linear regression models include those of Ermini and Hendry (2008) and Kobayashi and McAleer (1999); see also Ericsson (1982).

References

Antle, C.E., Bain, L.J.: A property of maximum likelihood estimators of location and scale parameters. SIAM Rev. **11**, 251–253 (1969)

Araujo, M.I. , Fernandes, M., Pereira, B.B.: Alternative procedures do discriminate non-nested multivariate linear regression models. Commun. Stat. Theory Meth. **34**, 2047–2062 (2005)

Arkonac, S.Z., Higgins, M.L.: A Monte Carlo Study of tests for non-nested models estimated by generalized method of moments. Commun. Stat. Simul. Comput. **24**, 745–763 (1995)

Atkinson, A.C.: A test for descriminating between models. Biometrika **56**, 337–347 (1969)

Atkinson, A.C.: A method for descriminating between models (with discussion). J. R. Stat. Soc. B 323–353 (1970)

Cox, D.R.: Tests of separate families of hypotheses. In: Proceedings 4th Berkeley Symposium in Mathematical Statistics and Probability, vol. 1, pp. 105–123, University of California Press (1961)

Cox, D.R.: Further results on test of separate families of hypotheses. J. R. Stat. Soc. **B**, 406–424 (1962)

Cox, D.R.: A return to an old paper: tests of separate families of hypotheses (with discussion). J. R. Stat. Soc. B **75**, 207–215 (2013)

Cox, D.R., Hinkley, D.V.: A note on the efficiency of least squares estimates. J. R. Stat. Soc. **71**, 284–289 (1968)
Cox, D.R., Hinkley, D.V.: Problems and solutions in theoretical statistics. In: Chapman and Hall (eds.) (1978)
Cribari-Neto, F., Lucena, S.E.F.: Nonnested hypothesis testing in the class of varying dispersion beta regressions. J. Appl. Stat. **42**, 967–985 (2015)
Dastoor, N.K.: A classical approach to Cox's test for non-nested hypotheses. J. Econ. **27**, 363–370 (1985)
Davidson, R., MacKinnon, J.G.: Several tests for model specification in the presence of alternative hypotheses. Econometrica **49**, 781–793 (1981)
Davidson, R., MacKinnon, J.G.: Some non-nested hypothesis tests and the relations among them. Rev. Econ. Stud. **59**, 551–565 (1982)
Davidson, R., MacKinnon, J.G.: Testing the specification of multivariate models in the presence of alternative hypotheses. J. Econ. **23**, 301–313 (1983)
Dey, A.K., Kundu, D.: Discriminating among the log-normal, Weibull and generalized exponential distributions. IEEE Trans. Reliab. **58**, 416–424 (2009)
Dey, A.K., Kundu, D.: Discriminating between the bivariate generalized exponential and bivariate Weibull distributions. Chil. J. Stat. **3**, 93–110 (2012a)
Dey, A.K., Kundu, D.: Discriminating between the Weibull and log-normal distribution for type-II censored data. Stat. J. Theoret. Appl. Stat. **46**, 197–214 (2012b)
Dumonceaux, R., Antle, C.E.: Discriminating between the lognormal and the Weibull distribution. Technometrics. **15**, 923–926 (1973)
Dumonceaux, R., Antle. C.E, Haas, G.: Likelihood ratio test for discriminating between two models with unknown location scale parameters. Technometrics. **15**, 19–27 (1973a)
Efron, B.: Comparing non-nested linear models. J. Am. Stat. Assoc. **79**, 791–803 (1984)
Epps, T.W., Singleton, K.J., Pulley, L.B.: A test of separate families of distributions based on the empirical moment generating function. Biometrika **69**, 391–399 (1982)
Ericsson, N.R.: Testing linear versus logarithmic regression models: a comment. Rev. Econ. Stud. **49**, 447–481 (1982)
Ermini, L., Hendry, D.F.: Log income vs. linear income: an application of encompassing principle. Oxford Bull. Econ. Stat. **70**, Supplement, 807–827 (2008)
Ferrari, S.L.P., Cribari-Neto, F.: Beta regression for modelling rates and proportions. J. Appl. Stat. **31**, 799–815 (2004)
Fine, J.P.: Comparing non-nested Cox models. Biometrika **89**, 635–647 (2002)
Fisher, G.R.: Tests for two separate regressions. J. Econ. **21**, 117–132 (1983)
Fisher, G.R., McAleer, M.: Alternative procedures and associated tests of significance for non-nested hypotheses. J. Econ. **16**, 103–119 (1981)
Gould, A., Lawless, J.F.: Consistency and efficiency of regression coefficient estimates in location-scale models. Biometrika **75**, 535–540 (1988)
Gupta, R.D., Kundu, D.: Discriminating between gamma and generalized exponential distributions. J. Stat. Comput. Simul. **74**, 107–121 (2004)
Hagemann, A.: A simple test for regression specification with non-nested alternatives. J. Econ. **166**, 247–254 (2012)
Hottelling, H.: The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. Ann. Math. Stat. **11**, 271–283 (1940)
Jackson, O.A.Y.: Some results on tests of separate families of hypotheses. Biometrika **55**, 355–363 (1968)
Jackson, O.A.Y.: Fitting a gamma or log-normal distribution to fibre-diameter on wool tops. Appl. Stat. **18**, 70–75 (1969)
Kent, J.T.: Robust properties of likelihood ratio test. Biometrika **69**, 19–27 (1982)
Kent, J.T.: The underlying structure of non-nested hypotheses tests. Biometrika **73**, 333–343 (1986)
Kobayashi, M., McAleer, M.: Analytical power comparisons of nested and nonnested tests for linear and loglinear regression models. Econ. Theory **15**, 99–113 (1999)

Kotz, S.: Normality vs. lognormality with applications. Commun. Stat. **1**, 113–132 (1973)
Kundu, D.: Discriminating between the normal and Laplace distributions. In: Balakrishnan, N., Nagaraja, H.N., Kannan, K. (eds.) Advances in Ranking and Selection, Multiple Comparisons and Reliability, pp. 65–85. Birkhouse (2005)
Kundu, D., Raqab, M.Z.: Discriminating between the generalized Rayleigh and log-normal distribution. Stat. J. Theoret. Appl. Stat. **41**, 505–516 (2007)
Lanot, G.: On the variance covariance matrix of the maximum likelihood estimator of discrete mixture. Working Paper Center for Economic Research, Keele University (KERP 2002/7) (2002)
McAleer, M.: The significance of testing empirical non-nested models. J. Econ. **67**, 149–171 (1995)
Oakes, D.: Direct calculation of information matrix via EM algorithm. J. R. Stat. Soc. B **61**, 479–482 (1999)
Otsu, T., Seo, M.H., Whang, Y.J.: Testing for non-nested conditional moment restrictions using unconditional empirical likelihood. J. Econ. **167**, 370–382 (2012)
Pereira, B.B.: Some results on tests of separate families of hypotheses. Ph.D. Thesis (Statistics), Imperial College, University of London (1976)
Pereira, B.B.: Empirical comparisons of some tests of separate families of hypotheses, Part II. Atas do 10 Simposio Brasileiro de Pesquisa Operacional, v.2 (1977a)
Pereira, B.B.: A note on the consistency and on finite sample comparisons of some tests of separate families of hypotheses. Biometrika **64**, 109–113 (1977b)
Pereira, B.B.: Tests of efficiencies of separate regression models. Biometrika, **65**, 319–327 (Amendment in Biometrika, **68**, 345, 1981) (1978)
Pereira, B.B.: Testes para discriminar entre as distribuições lognormal, gama e Weibull. Estadistica J. Inter-Am. Stat. Inst. **33**, 41–46 (1979)
Pereira, B.B.: Empirical comparisons of some tests of separate families of hypotheses. Metrika **25**, 219–234 (1981a)
Pereira, B.B.: Choice of a survival model for patients with a brain tumour. Metrika **28**, 53–61 (1981b)
Pereira, B.B.: On the choice of a Weibull model. Estadistica J. Inter-Am. Stat. Inst. **26**, 157–163 (1984)
Pereira, B.B.: Separate families of hypotheses. In: Armitage, P., Calton, T. (eds.) Encyclopedia of Biostatistics, 2nd edn, vol. 7, pp. 4881–4886. Wiley (2005)
Pereira, B.B.: Tests for discriminating separate or non-nested models. In: Lovic, M. (ed.) International Encyclopedia of Statistical Science, vol. 3, pp. 1592–1595. Springer (2010)
Pesaran, M.H.: On the general problem of model selection. Rev. Econ. Stud. **41**, 153–171 (1974)
Pesaran, M.H., Deaton, A.S.: Testing non-nested nonlinear regression models. Econometrica **46**, 677–694 (1978)
Pesaran, M.H.: Pitfalls on testing non-nested hypotheses by the Lagrange multiplier method. J. Econ. **17**, 323–331 (1981)
Pesaran, M.H.: On the comprehensive method of testing non-nested regression models. J. Econ. **18**, 263–274 (1982)
Pesaran, M.H.: Asymptotic power comparisons of tests of separate parametric families by Bahandur's approach. Biometrika **71**, 245–252 (1984)
Quandt, R.E.: A comparison of methods for testing nonnested hypotheses. Rev. Econ. Stat. **56**, 92–99 (1974)
Rao, C.R.: Large sample tests of statistical hypotheses concerning several parameters with applications to problem of estimation. Proc. Comb. Philos. Soc. **44**, 40–57 (1947)
Rojas, F.A.R.: Evaluation of test of separate hypotheses. Ph.D. Thesis (Operational Research), COPPE/UFRJ–Federal University of Rio de Janeiro (in Portuguese) (2001)
Rojas, F.A.R., Louzada-Neto, F., Pereira, B.B.: A note on some pitfalls on the sawyer test for descriminating between separate families of hypotheses. Braz. J. Prob. Stat. **22**, 85–88 (2008)
Sawyer, K.R.: Testing separate families of hyphotheses: an information criterion. J. R. Stat. Soc. B **45**, 89–99 (1983)
Sawyer, K.R.: Multiple hypotheses testing. J. R. Stat. Soc. B **46**, 419–424 (1984)

Shen, S.M.: A method for descriminating between models describing compositional data. Biometrika **69**, 587–595 (1982)
Silva, J.M.C.S.: A score test for non-nested hypotheses with applications to discrete data models. J. Appl. Econ. **16**, 577–591 (2001)
Slud, E.V.: Testing separate families of hypotheses using right censored data. Commun. Stat. Simul. Comput. **12**, 507–509 (1983)
Smith, R.J.: Non-nested tests for competing models estimated by generalized method of moments. Econometrica **60**, 973–980 (1992)
Terrel, G.R.: The gradient statistic. Comput. Sci. Stat. **34**, 206–215 (2002) (Symposium on the interface: Computing and Statistics, Montereal, Canada, April, 19, 2002)
Timm, N.H., Al-Subaihi, A.A.: Testing model specification in seemingly unrelated regression models. Commun. Stat. Theory Meth. **30**, 579–590 (2001)
Vuong, Q.H.: Likelihood ratio test for model selection and non-nested hypotheses. Econometrika. **52**, 307–333 (1989)
Wald, A.: Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Am. Math. Soc. **54**, 426–482 (1943)
Walker, M.D., Gehan, E.A., Haventhal, C.M., Morrel, H.A., Mahaley, M.S.: The evaluation Mithramycin (NSC-24599) in the treatment of anaplastic gliomas. In: Presented at the Fourth International Congress of Neurological Surgery, New York (1969)
White, H.: Regularty conditions for Cox's test of non-nested hypotheses. J. Econ. **19**, 301–318 (1982)
Wilks, S.S.: The large sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9**, 60–62 (1938)

Chapter 3
Bayesian Methods

Abstract Bayesian methods of model discrimination are discussed in this chapter. Alternative Bayes factors are presented for when improper priors are used and the usual Bayes factor cannot be specified. The concepts of imaginary training sample and minimal training samples and of partial, fractional, intrinsic and posterior Bayes factors are defined. Applications of these concepts to alternative (exponential, gamma, Weibull and lognormal) distributions and to systems of linear regressions are presented. Simulation results are used to compare the alternative Bayes factors. The Full Bayesian Significance Test (FBST) is also presented, with applications to a linear mixture model.

Keywords Alternative Bayes factors · Discrimination · Exponential distribution · FBST procedure · Gamma distribution · Improper prior · Linear mixture · Lognormal distribution · Predictive distribution · Weibull distribution

3.1 Introduction

In this chapter, the Bayesian approach to discriminating among separate models is studied.

To illustrate the Bayesian method for discriminating among separate models, consider the following (Fig. 3.1):

Example 3.1 (Pereira and Polpo 2014) Let $Y_n = y_1, ..., y_n$ be a set of points selected in the real plane as follows:

Our objective is to choose in which of two geometric figures, a circle or a square, these points are observed. To simplify the choice problem, we consider that both figures are centered at (0, 0). Hence, the two likelihoods that correspond to the circle and square are, respectively,

$$L_c(Y_n) \propto \mathbf{I}(\text{all points with booth coordinates} < \alpha)/\pi^n \alpha^{2n}, \alpha \geq D,$$
$$L_s(Y_n) \propto \mathbf{I}(\text{all points with booth coordinates} < \beta)/4^n \beta^{2n}, \beta \geq M,$$

B. de B. Pereira and C.A. de B. Pereira,
Model Choice in Nonnested Families, SpringerBriefs in Statistics,
DOI 10.1007/978-3-662-53736-7_3

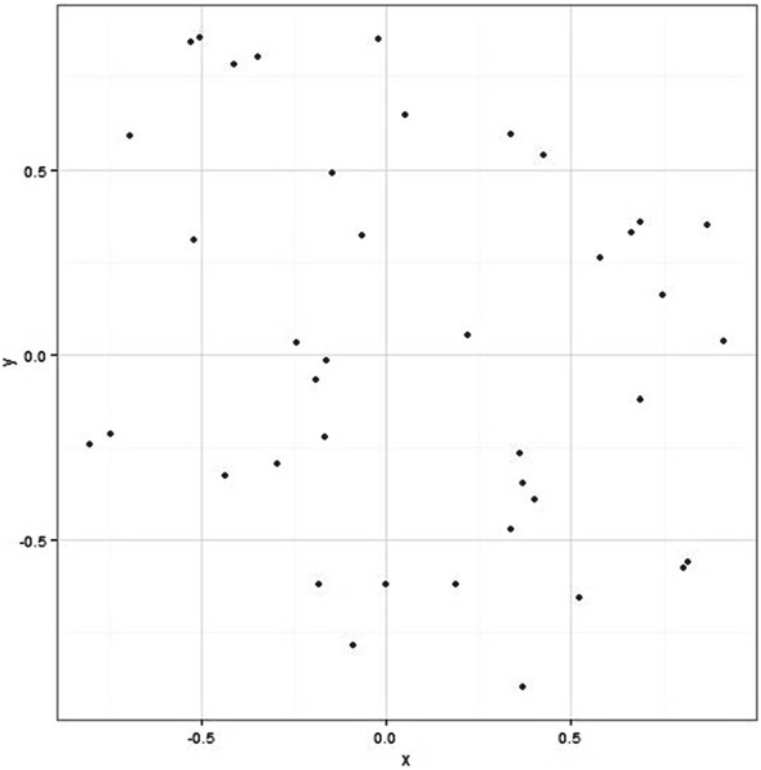


Fig. 3.1 Sample of 40 points in the plane

where α is the radius of the circle, 2β is the side length of the square, D is the largest of the distances from a point to the center of the circle, and M is the maximum value of the maximum of the two absolute values of the coordinates of each point.

The probability priors for the circle and square models are π_c and π_s, respectively, such that $\pi_c + \pi_s = 1$, and the priors for the parameters are $\pi_c(\alpha)$ and $\pi_s(\beta)$, respectively.

As a simplification, we consider these priors as being proportional to α^{-2} and β^{-2}, respectively, on the interval from 0.01 to infinity. Clearly, it would not be practical to consider zero as the lower limit because at least two sample points were obtained. The posterior odds ratio from (1.3) is

$$\frac{\pi_c}{\pi_s} B_{CS}(Y_n) = \frac{\pi_c}{\pi_s} \left(\frac{4}{\pi}\right)^n \left(\frac{M}{D}\right)^{2n+1}.$$

As the observed points are on the same surface, we consider $\pi_c = \pi_s = 1/2$; consequently, the posterior odds ratio is the Bayes factor. Through the Bayes estimation of α and β, we obtain the estimated circle and square for the two sets of samples described in Table 3.1.

Table 3.1 The two samples used for estimating the circle and square

Sample 1				Sample 2			
x_1	y_1	D_1	M_1	x_2	y_2	D_2	M_2
−0.181	−0.620	0.646	0.620	0.905	0.646	1.112	0.905
−0.434	−0.325	0.542	0.434	−0.501	−0.688	0.851	0.688
−0.087	−0.786	0.791	0.786	−0.058	−0.098	0.114	0.098
0.339	0.595	0.685	0.595	−0.161	−0.587	0.609	0.587
0.369	−0.347	0.506	0.369	−0.114	0.250	0.275	0.250
0.361	−0.265	0.448	0.361	−0.627	−0.488	0.795	0.627
0.188	−0.621	0.649	0.621	−0.334	−0.493	0.595	0.493
0.688	0.357	0.775	0.688	−0.822	−0.821	1.162	0.822
−0.166	−0.223	0.278	0.223	−0.100	0.408	0.420	0.408
−0.691	0.591	0.909	0.691	−0.805	0.384	0.892	0.805
−0.022	0.853	0.853	0.853	−0.514	−0.415	0.661	0.514
−0.502	0.858	0.994	0.858	−0.993	−0.370	1.059	0.993
−0.521	0.309	0.606	0.521	0.904	0.027	0.904	0.904
0.685	−0.123	0.696	0.685	−0.810	0.650	1.038	0.810
−0.410	0.783	0.884	0.783	−0.036	0.166	0.170	0.166
0.052	0.648	0.650	0.648	−0.600	−0.254	0.651	0.600
−0.802	−0.243	0.838	0.802	0.551	−0.408	0.685	0.551
−0.001	−0.621	0.621	0.621	−0.459	−0.623	0.774	0.623
−0.145	0.492	0.513	0.492	−0.492	0.438	0.659	0.492
−0.067	0.320	0.326	0.320	0.710	−0.845	1.103	0.845
−0.295	−0.293	0.416	0.295	0.336	0.612	0.698	0.612
−0.745	−0.213	0.774	0.745	−0.482	−0.694	0.845	0.694
0.749	0.161	0.767	0.749	−0.072	0.050	0.088	0.072
0.428	0.539	0.688	0.539	0.336	0.806	0.873	0.806
0.867	0.349	0.935	0.867	−0.638	0.546	0.840	0.638
0.401	−0.392	0.561	0.401	0.479	−0.293	0.561	0.479
0.337	−0.469	0.578	0.469	−0.330	−0.412	0.528	0.412
−0.527	0.842	0.994	0.842	−0.638	−0.118	0.649	0.638
0.663	0.330	0.741	0.663	0.776	−0.834	1.139	0.834
−0.348	0.805	0.877	0.805	0.935	0.722	1.181	0.935
0.220	0.053	0.226	0.220	0.432	−0.279	0.515	0.432
0.523	−0.658	0.840	0.658	0.661	−0.112	0.671	0.661
−0.161	−0.017	0.162	0.161	−0.601	−0.993	1.161	0.993
0.578	0.260	0.634	0.578	−0.826	−0.075	0.830	0.826
−0.244	0.031	0.246	0.244	0.981	0.772	1.248	0.981
0.817	−0.558	0.989	0.817	−0.122	−0.306	0.330	0.306

(continued)

Table 3.1 (continued)

Sample 1				Sample 2			
x_1	y_1	D_1	M_1	x_2	y_2	D_2	M_2
0.914	0.034	0.915	0.914	−0.550	−0.187	0.581	0.550
0.804	−0.576	0.989	0.804	−0.244	−0.390	0.460	0.390
0.371	−0.899	0.972	0.899	0.573	0.537	0.786	0.573
−0.189	−0.068	0.201	0.189	−0.820	0.577	1.002	0.820
Maximum		0.994	0.914	Maximum		1.248	0.993
Area		3.102	3.341	Area		4.895	3.945
LnOdds			15.674	LnOdds			−7.819

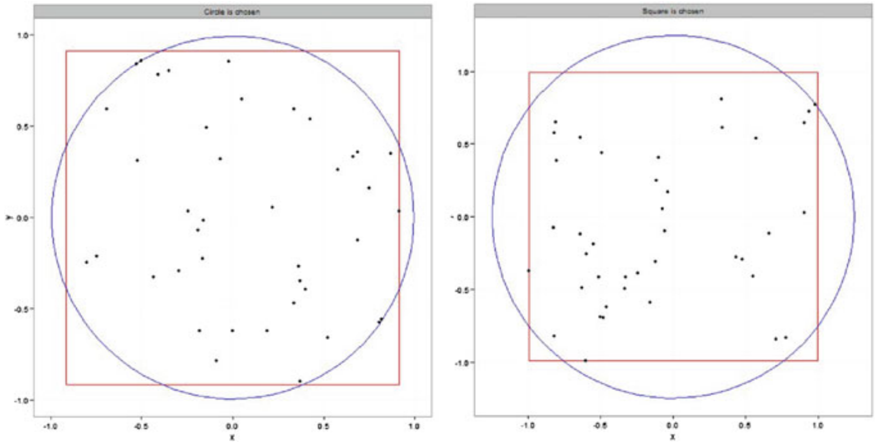


Fig. 3.2 Examples of samples with alternative choices: **a** the better choice is the circle; **b** the better choice is the square

The logarithm of the posterior odds ratio for sample 1 (Fig. 3.2a) is 15.674, indicating that the posterior probability of the circle is close to one, larger than the posterior probability of the square. The evidence from the data thus indicates that the candidate of choice is the circle, which has the smaller area. For sample 2, shown in Fig. 3.2b, the logarithm of the posterior odds ratio is −7.819, indicating that the better candidate is now the square.

Example 3.2 (Melo 2016) The gamma and lognormal distributions are two of the distributions that are most commonly used for positive random variables. Let us consider a Bayesian method for the linear mixture of these distributions. As suggested by Cox (1961), we can examine the estimates of the parameter mixture to decide on one of the models. From expressions (2.11) and (2.25), we can write

$$h_l(y, p, \alpha_1, \alpha_2, \delta_1, \delta_2) = pf_G(y, \delta_1, \delta_2) + (1 - p)f_L(y, \alpha_1, \alpha_2).$$

Let us now consider μ and σ^2 to be the true mean and variance, respectively, of the population. Therefore, for the lognormal distribution,

$$\mu = E(y, \alpha_1, \alpha_2) = e^{\alpha_1 + \alpha_2/2} \text{ and } \sigma^2 = V(y, \alpha_1, \alpha_2) = (e^{\alpha_1} - 1)e^{2\alpha_1 + \alpha_2},$$

and for the gamma distribution,

$$\mu = E(y, \delta_1, \delta_2) = \delta_1 \delta_2 \text{ and } \sigma^2 = V(y, \delta_1, \delta_2) = \delta_1^2 \delta_2^2.$$

Hence, there is a relationship between the parameters of the two models as described by μ and σ^2. The model parameters are now as follows: the connecting parameters are μ and σ^2, with p corresponding to the mixture. Initially, there were five parameters; now, there are only three parameters to be estimated.

Melo (2016) considered the following prior: the distributions for both μ and σ^2, the connecting parameters, are independent gamma distributions, both with a mean of one and a variance of 100. The prior for p is a beta prior with parameters (1, 1), the uniform distribution. For data on the survival times of 247 patients with cardiac insufficiency from a hospital in São Paulo, using the MCMC algorithm, the gamma distribution was found to be the preferred model, as shown in Table 3.2.

Figure 3.3 illustrates the fitting of the gamma-lognormal mixture, as estimated from the MCMC results, and the Kaplan–Meier estimates.

In general, applications of expressions (1.3)–(1.5) have two main limitations. First, the prior knowledge expressed by priors π_f and $\pi_f(\alpha)$ and by priors π_g and $\pi_g(\beta)$ must be coherent, as in Example 3.2. For instance, if the parameter spaces have different dimensions, in general, there is no simple relation between the parameters. Second, if the prior information is weak and an improper prior is applied, then the usual Bayes factor is not well defined. This problem arises when using the usual improper prior

$$\pi_f(\alpha) \propto h_f(\alpha) = c_f h_f(\alpha) \text{ and } \pi_g(\beta) \propto h_g(\beta) = c_g h_g(\beta), \quad (3.1)$$

where h_f and h_g denote functions whose integrals over the spaces of α and β diverge and c_f and c_g are unspecified normalizing constants.

Table 3.2 Mixture model—Bayes estimates

Parameter	Estimate	SD	LB 95 %	UB 95 %
p-Gamma	0.53	0.24	0.15	0.98
μ	13.59	1.06	11.60	15.72
σ^2	117.44	42.08	56.78	196.54

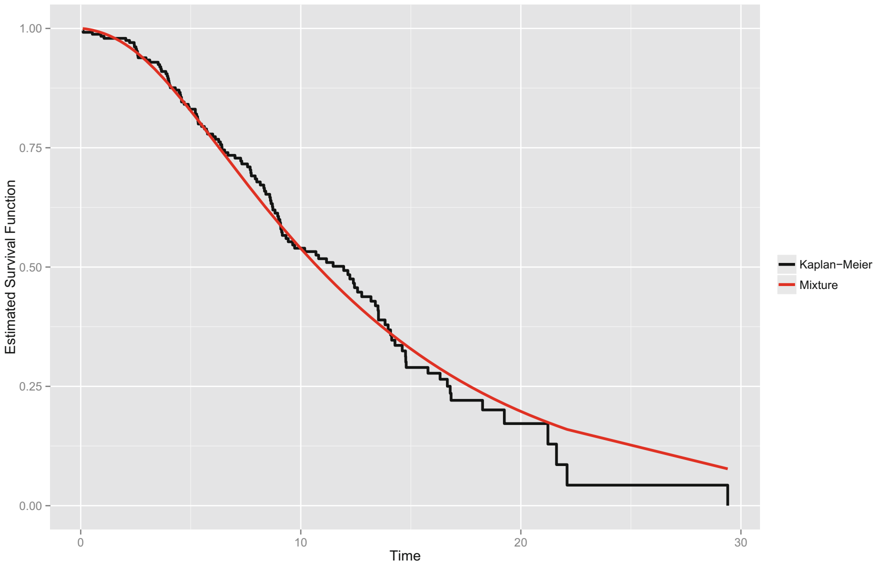


Fig. 3.3 Survival function estimated using the gamma-lognormal mixture and Kaplan–Meier estimates representing the observed data

If improper prior distributions are assigned to both models and we assume that $\pi_f = \pi_g$, then the posterior odds ratio (1.3) is

$$\frac{\pi_f}{\pi_g} B_{fg}(y) = B_{fg}(y) = \frac{q_f(y)}{q_g(y)} = \frac{\int f(y,\alpha)\pi_f(\alpha)d\alpha}{\int g(y,\beta)\pi_g(\beta)d\beta}$$
$$= \frac{c_f \int f(y,\alpha)h_f(\alpha)d\alpha}{c_g \int g(y,\beta)h_g(\beta)d\beta}. \tag{3.2}$$

The Bayes factor, which depends on c_f/c_g, is unspecified; $\int f(y,\alpha)h_f(\alpha)d\alpha$ and $\int g(y,\beta)h_g(\beta)d\beta$ are the predictive distributions under f and g, respectively, say m_f and m_g.

In the following Sect. 3.2, improper prior distributions are assumed, and to overcome the resulting difficulties, modified Bayes factors are described. Only basic definitions and examples are presented. Section 3.3 then presents the newly developed Full Bayesian Significance Test and its application to a linear mixture of models. References for discussions on and the properties of these procedures are briefly noted in Sect. 3.4.

3.2 Modified Bayes Factors

3.2.1 Imaginary Training Sample

To eliminate the indeterminacy of the Bayes factor with improper priors, Smith and Spiegelhalter (1980), and Spiegelhalter and Smith (1982) imagined a set of data y_0 for which a particular value is assigned to $B_{fg}(y_0)$, and so c_f/c_g are determined. Aitchinson (1978) and Pericchi (1984) also used these ideas to investigate some misleading behaviors of posterior probabilities. These procedures are related to the assignment of a certain kind of prior information and imply the rejection of the use of improper priors. Further critiques are presented in O'Hagan (1995).

3.2.2 Partial Bayes Factor (PBF)

An early solution to the indeterminacy of c_f/c_g was presented by Lempers (1971), who set aside part of the data to be combined with an improper prior distribution to produce a proper posterior distribution, which was then used to compute the Bayes factor from the remainder of the data.

Rust and Schmittlein (1985) also used the idea of training samples. In their Bayesian cross-validated likelihood method, the first subset of the sample is used to estimate the parameters, and Bayes' Theorem is then applied with these estimates to the second part of the sample.

A formal study of this idea of training samples appears in O'Hagan (1995). Consider the partitioning $y = (x, z)$ of the sample. From subsample x, one can obtain the proper posterior distributions $\pi_f(\alpha|x)$ and $\pi_g(\beta|x)$. With these as prior distributions, the remaining data z are then used to compute a Bayes factor:

$$B^p_{fg}(z|x) = \frac{q_f(z|x)}{q_g(z|x)} = \frac{\int \pi_f(\alpha|x) f(z, \alpha|x) d\alpha}{\int \pi_g(\beta|x) g(z, \beta|x) d\beta}. \tag{3.3}$$

Noticing that

$$q_f(z|x) = \frac{q_f(z.x)}{q_f(x)} = \frac{q_f(y)}{q_f(x)} \tag{3.4}$$

and $\pi_f(\alpha) = c_f h_f(\alpha)$, it follows that c_f can be removed.

The same is true for c_g.

It follows from (3.4) and the analogous relation for $q_g(z|x)$ that

$$B^p_{fg}(y) = B_{fg}(x) B_{fg}(z|x). \tag{3.5}$$

$B^p_{fg}(z|x)$ is referred to as a partial Bayes factor (PBF).

3.2.3 Fractional Bayes Factor (FBF)

Let $y = (x, z)$ be a sample of size n, and let x be a subsample of size m. To avoid the arbitrariness of choosing a particular subsample or to consider all possible subsamples of a given size, O'Hagan (1995) developed a simplified form of the partial Bayes factor. Let $b = m/n$. If both n and m are large, then the likelihoods $f(x, \alpha)$ and $g(x, \beta)$ based only on the training sample x approximate the full likelihoods $f(y, \alpha)$ and $g(y, \beta)$, respectively, both raised to the power b.

By analogy to Eqs. (3.3) and (3.4), the fractional Bayes factor (FBF) is defined as

$$B^b_{fg}(y) = q_f(b, y)/q_g(b, y), \tag{3.6}$$

where

$$q_f(b, y) = \frac{\int \pi_f(\alpha) f(y, \alpha) d\alpha}{\int \pi_f\{f(y, \alpha)\}^b d\alpha} \text{ and } q_g(b, y) = \frac{\int \pi_g(\alpha) g(y, \alpha) d\beta}{\int \pi_g\{g(y, \beta)\}^b d\beta}. \tag{3.7}$$

If $\pi_f(\alpha) = c_f h_f(\alpha)$ and $\pi_g(\beta) = c_g h_g(\beta)$, then the indeterminate constants c_f and c_g cancel out. O'Hagan (1995) showed that the FBF is consistent, provided that b shrinks to zero as n grows.

3.2.4 Intrinsic Bayes Factor (IBF)

In proposing another modified Bayes factor, Berger and Pericchi (1996) first defined a minimal training sample: x in the sample partition $y = (x, z)$ is minimal if the posteriors for α and β are proper and there is no subset of x that entails a proper posterior. There are usually many, say R, partitions that feature a minimal training sample. The intrinsic Bayes factor (IBF) of Berger and Pericchi is the geometric or arithmetic mean or the median of the partial Bayes factors $\left\{B^p_{fg}(z_r|x_r);\ r = 1, ..., R\right\}$ obtained from these R minimal training samples.

The geometric IBF is

$$B^{IG}(y) = \left\{\prod_{r=1}^{R} B^p_{fg}(z_r|x_r)\right\}^{1/R}, \tag{3.8}$$

the arithmetic IBF is

$$B^{IA}(y) = \frac{1}{R}\sum_{r=1}^{R} B_{fg}(z_r|x_r), \tag{3.9}$$

and the median IBF is

$$B^{IM}(y) = med\{B_{fg}(z_r|x_r);\ r = 1, ..., R\}. \tag{3.10}$$

Because all these versions are based on the PBF, the indeterminacy due to c_f/c_g disappears.

3.2.5 *Posterior Bayes Factor (POBF)*

Aitkin (1991) proposed the posterior Bayes factor (POBF), which compares the posterior means of the likelihood functions under H_f and H_g. More formally, the posterior densities under H_f and H_g are, respectively,

$$\pi_f(\alpha|y) = \frac{f(y,\alpha)\pi_f(\alpha)}{\int f(y,\alpha)\pi_f(\alpha)d\alpha} \text{ and } \pi_g(\beta|y) = \frac{g(y,\beta)\pi_g(\beta)}{\int f(y,\beta)\pi_g(\beta)d\beta}. \tag{3.11}$$

The posterior means of the likelihood functions are, respectively,

$$q_f^{PO}(y) = \int f(y,\alpha)\pi_f(\alpha|y)d\alpha = \frac{\int \{f(y,\alpha)\}^2 \pi_f(\alpha)d\alpha}{\int f(y,\alpha)\pi_f(\alpha)d\alpha}$$

and

$$q_g^{PO}(y) = \int g(y,\beta)\pi_g(\beta|y)d\beta = \frac{\int \{g(y,\beta)\}^2 \pi_g(\beta)d\beta}{\int f(y,\beta)\pi_g(\beta)d\beta}. \tag{3.12}$$

The POBF is defined as

$$B_{fg}^{PO}(y) = q_f^{PO}(y)/q_g^{PO}(y). \tag{3.13}$$

3.2.6 *Applications*

Before we present some examples of the use of the modified Bayes factors, it is important to present Jeffreys' rule (Kass and Raftery 1995), which provides the background for interpreting the Bayes factors (Table 3.3).

Example 3.3 (Araujo and Pereira 2007) Consider a single random sample $y = (y_1, ..., y_n)$. Three models are considered for the data y: $LN - lognormal\ LN(\mu, \sigma)$, $W - Weibull\ W(\beta_1, \beta_2)$ and $G - Gamma\ G(r, \lambda)$.

Table 3.3 Jeffreys' rule for Bayes factors

$2 \ln B_{fg}$	B_{fg}	Evidence against H_g
0–2	1–3	Not worth more than a bare mention
2–6	3–20	Substantial
6–10	20–150	Strong
>10	>150	Decisive

Table 3.4 Distribution specifications

Densities
$p_L(y\|\mu,\sigma) = \frac{1}{y\sqrt{2\pi}\sigma} \exp\left[\frac{-1}{2\sigma^2}(\ln(y)-\mu)^2\right], \sigma > 0, \quad -\infty < \mu < \infty, \quad y > 0$
$p_W(y\|\beta_1,\beta_2) = \frac{\beta_2}{\beta_1^{\beta_2}} y^{\beta_2 - 1} \exp\left[-\left(\frac{y}{\beta_1}\right)^{\beta_2}\right], \beta_1 > 0, \quad \beta_2 > 0, \quad y > 0$
$p_G(y\|r,\lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y}, \qquad r, \lambda, y > 0$
Likelihoods
$L_L(\mu,\sigma; y) = \frac{1}{\prod_{i=i}^n y_i \left(\sqrt{2\pi}\sigma\right)^n} \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^n (\ln(y_i)-\mu)^2\right]$
$L_W(\beta_1,\beta_2; y) = \frac{\beta_2^n}{\beta_1^{n\beta_2}} \prod_{i=1}^n y_i^{\beta_2 - 1} \exp\left[-\frac{1}{\beta_1^{\beta_2}}\sum_{i=1}^n y_i^{\beta_2}\right], \qquad y > 0$
$L_G(r,\lambda; y) = \frac{\lambda^{nr}}{[\Gamma(r)]^n} \prod_{i=1}^n y_i^{r-1} \exp\left\{-\lambda \sum_{i=1}^n y_i\right\}$
Priors
$\pi(\mu,\sigma) \propto \frac{1}{\sigma}$
$\pi(\beta_1,\beta_2) \propto \frac{1}{\beta_1\beta_2}$
$\pi(r,\lambda) \propto \frac{1}{\lambda\sqrt{r}}$
Predictives
$m_L(y) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\left(\prod_{i=i}^n y_i\right)\pi^{(n-1)/2} 2\sqrt{n}\left[\sum_{i=1}^n (\ln(y_i)-\bar{y}_L)^2\right]^{n-1}}, \quad \text{where } \bar{y}_1 = \frac{1}{n}\sum_{i=1}^n \ln y_i$
$m_W(y) = (n-1)! \int_0^\infty \frac{\beta_2^{n-2}}{\left(\sum_{i=1}^n y_i^{\beta_2}\right)^n} \left(\prod_{i=1}^n y_i^{\beta_2 - 1}\right) d\beta_2$
$m_G(y) = \frac{1}{\prod_{i=1}^n y_i} \int_0^\infty \left[\frac{\prod_{i=1}^n y_i}{(\sum_{i=1}^n y_i)^n}\right]^r \frac{\Gamma(nr)}{[\Gamma(r)]^n \sqrt{r}} dr$

Tables 3.4 and 3.5 present the formulae for computing the modified Bayes factors. The final expressions for the modified Bayes factors are obtained from these tables. For example, to discriminate between $LN(\mu,\sigma)$ and $W(\beta_1,\beta_2)$, we can calculate the FBF and POBF:

Table 3.5 Expressions for the modified Bayes factors

Minimal Sample Predictive
$mq_L(x^{(l)}) = \frac{1}{2x_i x_j \left\vert \ln\left(\frac{x_i}{x_j}\right)\right\vert}$
$m_W(x^{(l)}) = \int_0^\infty \frac{(x_i x_j)^{\beta_2 - 1}}{\left(x_i^{\beta_2} + x_j^{\beta_2}\right)^2} d\beta_2 = \frac{1}{2x_i x_j \ln(x_i/x_j)}$
$m_G\left(x^{(l)}\right) = \frac{1}{x_i x_j} \int_0^\infty \left[\frac{x_i x_j}{(x_i + x_j)^2}\right]^r \frac{\Gamma(2r)}{[\Gamma(r)]^2 \sqrt{r}} dr$
Denominator of $q_i(b, y)$ of the PBF
$fracL = \int_0^\infty \int_{-\infty}^\infty [L_L(y)]^b \frac{1}{\sigma} d\mu d\sigma = \frac{\Gamma(\frac{bn-1}{2})}{2\sqrt{nb}(\prod_{i=i}^n y_i)^b \pi^{(bn-1)/2} \sqrt{\left[b \sum_{i=1}^n (\log(y_i) - \bar{y}_L)^2\right]^{bn-1}}}$
$fracW = \int_0^\infty \int_0^\infty [L_W(y)]^b \frac{1}{\beta_1 \beta_2} d\beta_2 d\beta_1 = (nb-1)! \int_0^\infty \frac{\beta_2^{nb-2}}{\left(b \sum_{i=1}^n y_i^{\beta_2}\right)^{nb}} \left[\prod_{i=1}^n y_i^{b(\beta_2 - 1)}\right] d\beta_2$
$fracG = \int_0^\infty \int_0^\infty \frac{\lambda^{bnr}}{[\Gamma(r)]^{bn}} \left[\prod_{i=1}^n y_i\right]^{b(r-1)} e^{-b\lambda \sum_{i=1}^n y_i} \frac{1}{\lambda\sqrt{r}} d\lambda dr$ $= \int_0^\infty \frac{(\prod_{i=1}^n y_i)^{b(r-1)}}{[\Gamma(r)]^{bn}\sqrt{r}} \frac{\Gamma(bnr)}{(\sum_{i=1}^n y_i)^{bnr}} dr$
Numerator of $q_i^{PO}(y)$ of the BF (mean posterior likelihood)
$m_{Lpost}(y) = \int_0^\infty \int_{-\infty}^\infty [L_L(y)]^2 \frac{1}{\sigma} d\mu d\sigma = \frac{\Gamma(\frac{2n-1}{2})}{(\prod_{i=i}^n y_i)^2 \pi^{(2n-1)/2} 2\sqrt{2n} \sqrt{\left[2 \sum_{i=1}^n (\ln(y_i) - \bar{y}_L)^2\right]^{2n-1}}}$
$m_{Wpost}(y) = \int_0^\infty \int_0^\infty [L_W(y)]^2 \frac{1}{\beta_1 \beta_2} d\beta_2 d\beta_1 = (2n-1)! \int_0^\infty \frac{\beta_2^{2n-2}}{\left(2 \sum_{i=1}^n y_i^{\beta_2}\right)^{2n}} \left[\prod_{i=1}^n y_i^{2\beta_2 - 2}\right] d\beta_2$
$m_{Gpost} = \int_0^\infty \int_0^\infty \frac{\lambda^{2nr}}{[\Gamma(r)]^{2n}} \left[\prod_{i=1}^n y_i\right]^{2(r-1)} e^{-2\lambda \sum_{i=1}^n y_i} \frac{1}{\lambda\sqrt{r}} d\lambda dr$ $= \int_0^\infty \frac{[\prod_{i=1}^n y_i]^{2(r-1)}}{[\Gamma(r)]^{2n}\sqrt{r}} \int_0^\infty \lambda^{2nr-1} e^{-2\lambda \sum_{i=1}^n y_i} d\lambda dr = \int_0^\infty \frac{(\prod_{i=1}^n y_i)^{2(r-1)}}{[\Gamma(r)]^{2n}\sqrt{r}} \frac{\Gamma(2nr)}{(\sum_{i=1}^n y_i)^{2nr}} dr$

$$FBF = B_{LW}^{(b)}(y) = \left(\frac{m_L(y)}{f_{racL}}\right) \Big/ \left(\frac{m_W(y)}{f_{racW}}\right)$$
$$\text{and} \tag{3.14}$$
$$POBF = B_{LW}^{PO}(y) = \left(\frac{m_{LPOS}(y)}{m_L(y)}\right) \Big/ \left(\frac{m_{WPOS}(y)}{m_W(y)}\right).$$

The IBF is obtained by computing the predictive distribution from the data z (y without y_i, y_j) via numerical integration, with priors $m_L(x_\ell)$ and $m_W(x_\ell)$ and likelihoods $L_L(\mu, \sigma, z^L)$ and $L_W(\mu, \sigma, z^W)$. The IBF is obtained using all possible y_i and y_j $(i \neq j)$.

Example 3.4 Araujo (1998), Araujo and Pereira (2007b) generated simulation results for the alternative modified Bayes factors to discriminate L versus W, L versus G, and G versus W as well as for the exponential distribution $E(\lambda)$. A total of 100 samples were used for each size n. A typical result for the lognormal versus Weibull distributions can be seen in Fig. 3.4 and Tables 3.6 and 3.7.

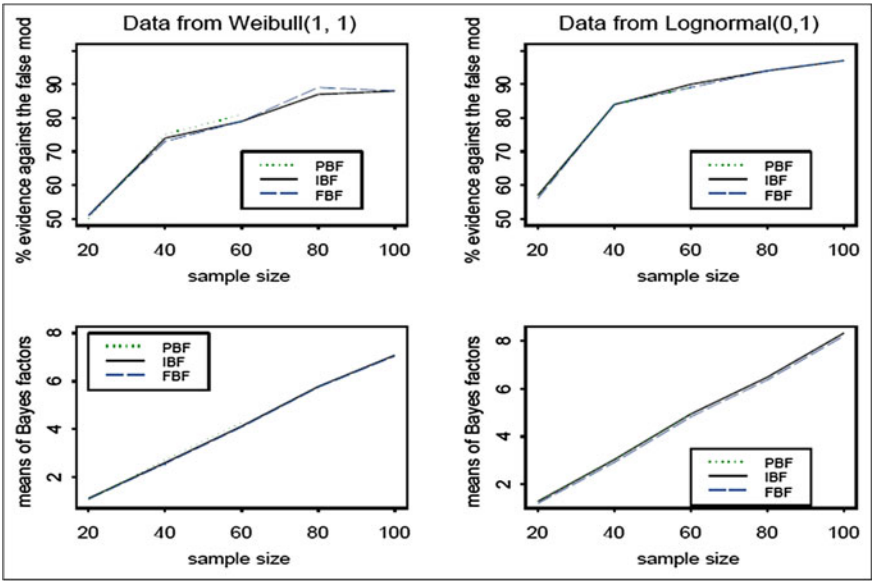


Fig. 3.4 Weibull versus lognormal

Table 3.6 Bayes factor LW, Data: L(0,1)

	Factor	min	median	max	mean	sd	$2 \ln FB > 2$ (%)	$2 \ln FB < 0$ (%)
$n = 20$	FBI	−2.085	1.290	6.511	1.28	1.438	57	16
	FBP	−2.263	1.305	6.758	1.291	1.505	57	16
	FBF	−2.003	1.198	6.125	1.187	1.358	56	16
$n = 40$	FBI	−6.194	2.878	13.220	3.033	2.496	84	7
	FBP	−6.414	2.892	13.440	3.046	2.550	84	7
	FBF	−6.061	2.776	12.830	2.925	2.427	84	7
$n = 60$	FBI	−4.703	5.051	15.170	4.940	3.325	90	8
	FBP	−4.841	5.075	15.310	4.953	3.370	89	8
	FBF	−4.649	14.86	4.823	4.823	3.264	89	8

They concluded that apart from the difficulties presented in the discussion provided by Aitkin (1991), the POBF should not be recommended because of the computational problems of instability and precision that arise with increasing n.

The behaviors of the IBF and FBF are similar, and the FBF requires less computational effort.

Note that because the lognormal and Weibull densities are in the location-scale form, the Bayes factors B_{WL} and B_{LW} are invariant with respect to the parameter values.

Table 3.7 Bayes factor WL, Data: W(1, 1)

	Factor	min	median	max	mean	sd	$2\ln FB > 2$ (%)	$2\ln FB < 0$ (%)
$n = 20$	FBI	−4.012	1.051	5.769	1.128	1.715	51	26
	FBP	−4.294	0.990	5.994	1.078	1.809	50	31
	FBF	−3.765	1.1016	5.494	1.091	1.624	51	26
$n = 40$	FBI	−4.572	2.454	13.890	2.606	2.820	74	18
	FBP	−4.597	2.562	14.33	2.728	2.894	75	18
	FBF	−4.413	2.412	13.570	2.564	2.747	73	18
$n = 60$	FBI	−3.978	3.973	14.940	4.132	3.625	79	12
	FBP	−3.927	4.115	15.230	4.285	3.670	81	11
	FBF	−3.882	3.930	14.710	4.088	3.564	79	12

Example 3.5 (Araujo et al. 2005, 2007) This example extends the results of Aitkin (1991), O'Hagan (1995) and Berger and Pericchi (1996) to the context of multivariate regressions.

Consider two separate multivariate linear regression models $H_0 : Y = XB_0 + U_0$ and $H_1 : Y = ZB_1 + U_1$, where Y is an $n \times m$ matrix of regressands, X and Z are, respectively, $n \times p$ and $n \times q$ matrices of regressors, and B_0 and B_1 are, respectively, $p \times m$ and $q \times m$ matrices of parameters. The error terms U_0 and U_1 have rows that are iid as normal random vectors with mean zero and identity covariance matrices Σ_0 and Σ_1, respectively. We also assume that X and Z are of full rank, with $n \geq m + p$ and $n \geq m + q$. It thus follows that $U_0 \sim \mathcal{N}(0, I_n \otimes \Sigma_0)$ and $U_1 \sim \mathcal{N}(0, I_n \otimes \Sigma_1)$, whereas $Y \sim \mathcal{N}(XB_0, I_n \otimes \Sigma_0)$ under H_0 and $Y \sim \mathcal{N}(ZB_1, I_n \otimes \Sigma_1)$ under H_1.

The matrices of regressors X and Z are fixed and nonnested in the sense that it is not possible to obtain the columns of X from the columns of Z, and vice versa. We further assume that the matrices $\Sigma_{X'X} \equiv \lim_{n\to\infty} \frac{1}{n} X'X$ and $\Sigma_{Z'Z} \equiv \lim_{n\to\infty} \frac{1}{n} Z'Z$ are nonsingular and that $\Sigma_{X'Z} \equiv \lim_{n\to\infty} \frac{1}{n} X'Z$ is a nonzero matrix. These assumptions ensure that the maximum likelihood estimators $\hat{B}_0 = (X'X)^{-1}X'Y$ and $\hat{B}_1 = (Z'Z)^{-1}Z'Y$ are consistent under H_0 and H_1, respectively.

The posterior odds ratio for H_0 against H_1 is $(\pi_0/\pi_1)\, B_{01}$. Suppose that one uses improper priors for the parameters such that $\pi_0(\alpha_0)$ and $\pi_1(\alpha_1)$ are proportional to the constants K_0 and K_1, respectively. Then, the Bayes factor B_{01} is proportional to K_0/K_1 and is not well defined. For the multivariate regression models, Jeffreys' diffuse prior is given by

$$\pi_0(\alpha_0) = \pi_0(B_0)\, \pi_0(\Sigma_0) = K_0\, |\Sigma_0|^{-\frac{m+1}{2}}, \tag{3.15}$$

leading to the following predictive distribution under the null hypothesis

$$q_0(Y) = \pi^{\frac{m(2n-2p-m+1)}{4}}\, K_0\, |X'X|^{-m/2}\, |S_0|^{-\frac{n-p}{2}} \prod_{s=1}^{m} \Gamma\left(\frac{n-p-s+1}{2}\right), \tag{3.16}$$

where $S_0 \equiv \left(Y - X\hat{B}_0\right)' \left(Y - X\hat{B}_0\right)$. A similar expression holds for the alternative model $Y = ZB_1 + U_1$. The resulting Bayes factor is

$$B_{01}(Y) = \pi^{m(p-q)/2} \frac{K_0}{K_1} \left(\frac{|Z'Z|}{|X'X|}\right)^{m/2} \frac{|S_1|^{(n-q)/2}}{|S_0|^{(n-p)/2}} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{n-p-s-1}{2}\right)}{\Gamma\left(\frac{n-q-s-1}{2}\right)}, \tag{3.17}$$

where $S_1 \equiv (Y - Z\hat{B}_1)'(Y - Z\hat{B}_1)$. It is clear from (3.17) that the Bayes factor is not well defined because it depends on the unknown ratio K_0/K_1.

From (3.16) and (3.17), it is now possible to derive the alternative Bayes factors. For instance, the POBF $B_{01}^P(Y)$ of Aitkin (1991) is found from the ratio between

$$q_0^P(Y) = (2\sqrt{\pi})^{-mn} |S_0|^{-n/2} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{2n-p-s+1}{2}\right)}{\Gamma\left(\frac{n-p-s+1}{2}\right)}$$

and

$$q_1^P(Y) = (2\sqrt{\pi})^{-mn} |S_1|^{-n/2} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{2n-q-s+1}{2}\right)}{\Gamma\left(\frac{n-q-s+1}{2}\right)}.$$

It therefore follows that

$$B_{01}^P(Y) = \left(\frac{|S_1|}{|S_0|}\right)^{n/2} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{2n-p-s+1}{2}\right)\Gamma\left(\frac{n-q-s+1}{2}\right)}{\Gamma\left(\frac{2n-q-s+1}{2}\right)\Gamma\left(\frac{n-p-s+1}{2}\right)}. \tag{3.18}$$

The arithmetic IBF of Berger and Pericchi (1996) becomes

$$B_{01}^{IA}(Y) = \frac{1}{R}\sum_{r=1}^{R} \frac{B_{01}(Y)}{B_{01}(Y_{(r)})} = B_{01}(Y)\,\frac{1}{R}\sum_{r=1}^{R} B_{10}(Y_{(r)}),$$

where $Y_{(r)}$ is a minimal training sample with design matrices $X_{(r)}$ and $Z_{(r)}$ under H_0 and H_1, respectively. By definition, $Y_{(r)}$ is a matrix such that both $X'_{(r)}X_{(r)}$ and $Z'_{(r)}Z_{(r)}$ are nonsingular. It is of $\bar{n} \times m$ dimensions, where $\bar{n} = \lceil (m+1)/2 \rceil + \max(p, q)$ and $\lceil \cdot \rceil$ returns the smallest integer greater than its argument. From (3.17), it follows that

$$B_{01}^{IA}(Y) = \left(\frac{|Z'Z|}{|X'X|}\right)^{m/2} \frac{|S_1|^{\frac{n-q}{2}}}{|S_0|^{\frac{n-p}{2}}} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{n-p-s+1}{2}\right)\Gamma\left(\frac{\bar{n}-q-s+1}{2}\right)}{\Gamma\left(\frac{n-q-s+1}{2}\right)\Gamma\left(\frac{\bar{n}-p-s+1}{2}\right)}$$
$$\times \frac{1}{R}\sum_{r=1}^{R}\left(\frac{|X'_{(r)}X_{(r)}|}{|Z'_{(r)}Z_{(r)}|}\right)^{m/2} \frac{|S_{0(r)}|^{(\bar{n}-p)/2}}{|S_{1(r)}|^{(\bar{n}-q)/2}}, \quad (3.19)$$

where $S_{j(r)}$ is analogous to S_j for the r-th minimal training set ($j = 0, 1$).
Finally, the FBF of O'Hagan (1995) is found from the ratio between

$$q_0^{[b]}(Y) = \pi^{mn(1-b)/2} b^{mnb/2} |S_0|^{-n(1-b)/2} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{n-p-s+1}{2}\right)}{\Gamma\left(\frac{nb-p-s+1}{2}\right)}$$

and

$$q_1^{[b]}(Y) = \pi^{mn(1-b)/2} b^{mnb/2} |S_1|^{-n(1-b)/2} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{n-q-s+1}{2}\right)}{\Gamma\left(\frac{nb-q-s+1}{2}\right)}. \quad (3.20)$$

Thus, it holds that

$$B_{01}^{[b])}(Y) = \left(\frac{|S_1|}{|S_0|}\right)^{n(1-b)/2} \prod_{s=1}^{m} \frac{\Gamma\left(\frac{n-p-s+1}{2}\right)\Gamma\left(\frac{nb-q-s+1}{2}\right)}{\Gamma\left(\frac{n-q-s+1}{2}\right)\Gamma\left(\frac{nb-p-s+1}{2}\right)}. \quad (3.21)$$

Example 3.6 (Araujo and Pereira 2001a) Inflation in Brazil in the post-war period has been discussed by Barbosa (1983). There are two main schools of thought on inflation during the 1950s. Monetarists consider the exaggerated growth of the money supply to be the main cause of inflation. Structuralists argue that inflation was generated within the economic system through changes in relative prices resulting from economic growth. In this sense, inflation would originate from monetary policy, which is passive and accommodates the variations in the nominal income of the economy. Schematically, we can represent these two perspectives in the following ways:

Monetarism

$$Deficit\ Spending \Longrightarrow Money\ growth \Longrightarrow Inflation$$

Structuralism

$$Shortage\ of\ key\ goods \Longrightarrow Inflation \Longleftarrow Struggles\ between\ social\ groups$$

Table 3.8 Matrices of least-squares estimates

Monetarism: $\hat{B}_0$			Structuralism: $\hat{B}_1$		
	p_t	h_t		p_t	h_t
p_{t-1}	0.3232	0.1429	p_{t-1}	0.8262	0.092
h_{t-1}	−0.2886	0.7570	h_{t-1}	−0.1872	0.7728
μ_t	0.8674	−0.1526	$S_{m,t}$	0.0716	−0.0158
$D\bar{y}_t$	−1.5073	0.2955	DZ_t	0.0043	−0.0003
			1	−0.3974	−1.1388

The models can be written as follows:

MONETARISM

$$[\,p_t\ h_t\,] = \left[\,p_{t-1}\ h_{t-1}\ \mu_{t-1}\ D\bar{y}_t\,\right] \begin{bmatrix} \alpha_1 - \beta\alpha_2 & 1 \\ -\beta\alpha_1 & \alpha_1 \\ \beta & -1 \\ \beta\alpha_1 & \alpha_1 \end{bmatrix} \phi + \varepsilon, \tag{3.22}$$

where $1949 \le t \le 1980$; $\phi = 1/[\alpha_1 - \beta(1-\alpha_2)]$; $D\bar{y}_t = Dy_t + h_t + h_{t-1}$; and for year t, $p_t = inflation\ rate$, $h_t = idle\ capacity$, $\mu_t = rate\ money\ of\ growth$, $D\bar{y}_t = potential\ product\ rate\ of\ growth$, and $Dy_t = real\ product\ rate\ of\ growth$.

STRUCTURALISM

$$[\,p_t\ h_t\,] = \left[\,p_{t-1}\ h_{t-1}\ S_{m,t}\ DZ_t\ 1\,\right] \begin{bmatrix} \beta_{11} + \gamma_{12} & \gamma_{12}\beta_{11} + \beta_{21} \\ \beta_{12} + \gamma_1 2 & \gamma_{12}\beta_{12} + 1 \\ \beta_{13} & \beta_{13}\gamma_{21} \\ \gamma_{12}\beta_2 3 & \beta_{23} \\ \beta_{10} + \gamma_{12}\beta_{20} & \beta_{20} + \beta_{10} \end{bmatrix} \varphi + \varepsilon, \tag{3.23}$$

where $\varphi = 1/[1 - \gamma_{12}\gamma_2 1]$ and for year t, $S_{m,t} = minimum\ wage$, $DZ_t = budget\ deficit$, $p_t = inflation\ rate$, and $h_t = idle\ capacity$.

Table 3.8 presents the parameter estimates for the models $\hat{B}_0 = (X'X)^{-1}X'Y$ and $\hat{B}_1 = (Z'Z)^{-1}Z'Y$:

From the results of Exercise 3, the modified Bayes factors are as follows:

(a) FBF:

$$B^{[b]}_{mon\times est}(Y) = \left(\frac{|S_{est}|}{|S_{mon}|}\right)^{11,5} \frac{\Gamma(14)\Gamma(2)\Gamma(13,5)\Gamma(1,5)}{\Gamma(13,5)\Gamma(2,5)\Gamma(13)\Gamma(2)}$$

with $2\log B^{[b]}_{mon\times est}(Y) = 7,121$.

(b) POBF:

$$B^{P}_{mon\times est}(Y) = \left(\frac{|S_{est}|}{|S_{mon}|}\right)^{16} \frac{\Gamma(30)\Gamma(13,5)\Gamma(29,5)\Gamma(13)}{\Gamma(29,5)\Gamma(14)\Gamma(29)\Gamma(13,5)}$$

with $2\log B^{P}_{mon\times est}(Y) = 5,028$.

(c) From the time series data, we have 26 training samples and $\bar{n}$ such that the matrices $Z^{'}_{(e)}$ and $Z_{(e)}$ are nonsingular ($\bar{n} = 7$).

$$B^{IA}_{mon\times est}(Y) = \frac{|Z^{'}_{est}Z_{est}||S_{est}|^{13,5}\Gamma(14)\Gamma(2)\Gamma(13,5)\Gamma(1,5)}{|Z^{'}_{mon}Z_{mon}||S_{mon}|^{14}\Gamma(13,5)\Gamma(2,5)\Gamma(13)\Gamma(2)} \times \frac{1}{24}\sum_{l=1}^{24} \frac{|Z^{'}_{mon}(l)Z_{mon}(l)||S_{mon}(l)|^{2,5}}{|Z^{'}_{est}(l)Z_{est}(l)||S_{est}(l)|^{2}}$$

with $2\log B^{IA}_{mon\times est}(Y) = 27,172$.

For the models described, the three modified Bayes factors indicate that the monetarist model is the preferred explanation of inflation and idle capacity in the Brazilian post-war period.

3.3 Full Bayesian Significance Test (FBST)

The FBST of Pereira and Stern (1999), which is reviewed in Pereira et al. (2008), is a Bayesian version of significance testing as considered by Cox (1977) and Kempthorne (1976). The frequentist method of significance testing is a procedure for measuring the consistency of a set of data with the null hypothesis. The basis of the test is an ordering of the sample space according to increasing inconsistency with the hypothesis. The index used to measure this inconsistency is the calibrated p-value. By contrast, the basis for the Bayesian method is an index known as the e-value (where e stands for evidence), which measures the inconsistency of the hypothesis using several parameter points together with the posterior densities.

First, let us consider a real parameter ω, a point in the parameter space $\Omega \subset \Re$, and an observation y of the random variable Y. A frequentist looks for the set $I \in \Re$ of sample points that are at least as inconsistent with the hypothesis as y is. A Bayesian looks for the tangential set $T \in \Omega$ (Pereira et al. 2008), which is a set of parameter points that are more consistent with the observed y than the hypothesis is. An example of a sharp hypothesis in a parameter space of the real line is of the type $\mathbf{H} : \omega = \omega_0$. The evidence value in favor of $\mathbf{H}$ for a frequentist is the usual p-value, $P(Y \in I|\omega_0)$, whereas for a Bayesian, the evidence in favor of $\mathbf{H}$ is the e-value, $ev = 1 - \Pr(\omega \in T|y)$.

In the general case of multiple parameters, $\Omega \subset \Re^k$, let the posterior distribution for ω given y be denoted by $q(\omega|y) \propto \pi(\omega)L(y, \omega)$, where $\pi(\omega)$ is the prior probability density of ω and $L(y, \omega)$ is the likelihood function. In this case, a sharp hypothesis is of the type $\mathbf{H} : \omega \in \Omega_H \subset \Omega$, where Ω_H is a subspace of smaller dimension than Ω. Letting $\sup_H$ denote the supremum of Ω_H, we define the general Bayesian evidence and the tangential set as follows:

$$q^* = \sup_H q(\omega|y) \text{ and } T = \{\omega : q(\omega|y) > q^*\}. \tag{3.24}$$

The Bayesian evidence value against **H** is the posterior probability of T,

$$\overline{ev} = \Pr(\omega \in T|y) = \int_T q(\omega|y)d\omega; \text{ consequently, } ev = 1 - \overline{ev}. \tag{3.25}$$

It is important to note that evidence that favors **H** is not evidence against the alternative **A** because it is not a sharp hypothesis. This interpretation also holds for p-values in the frequentist paradigm. As in Pereira et al. (2008), we would like to point out that this Bayesian significance index uses only the posterior distribution, with no need for additional artifacts such as the inclusion of positive prior probabilities for the hypotheses or the elimination of nuisance parameters. In fact, it is not recommended to consider the construction of tangential sets in marginal distributions of the parameters of evidence. We should not abandon the original parameter space in its full dimensionality, without any complication due to the dimensionality of either the parameter or sample spaces. If we believe that there is no need for the use of prior information and that the integral of the likelihood is finite, then the normalized likelihood can serve as the posterior probability density: the measure of consistency of the hypothesis with the observed data is subject to no interference from prior knowledge. The computation of the e-values does not require asymptotic methods, and the only technical tools needed are numerical optimization and integration methods.

Example 3.7 (Example 3.2 cont.) For the hypothesis $\mathbf{H} : p = 0$, corresponding to rejection of the gamma model, we obtain an e-value of 0.002, which favors the gamma model. However, for the hypothesis $\mathbf{H} : p = 1$, we obtain an e-value of 0.8, which corresponds to not rejecting H with a corresponding p-value of 0.2. These p-values follow from Diniz et al. (2012). We conclude this section with a test of $\mathbf{H} : p = 0.5$, which yields a p-value of 0.99 with a corresponding p-value of 0.77 in favor of the mixture (see Table 3.9 and Fig. 3.3).

Table 3.9 Hypothesis testing for the mixture parameters of the gamma and lognormal models

Hypothesis	e-value	p-value
$p = 0$	0.002	0.00004
$p = 1$	0.80	0.20
$p = 0.5$	0.99	0.77

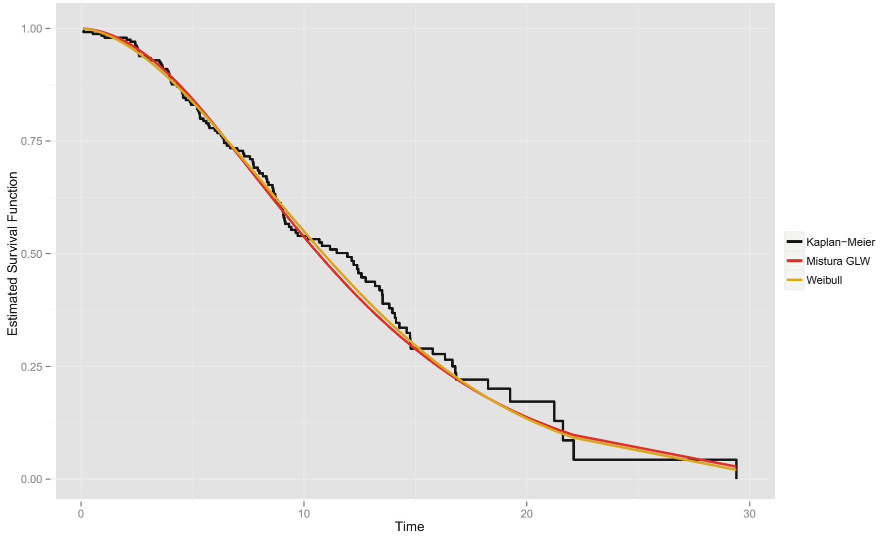


Fig. 3.5 Comparison of the Weibull and GLW survival estimates, with the Kaplan–Meier estimates representing the observed data

Example 3.8 The authors applied the following linear mixture of the models in (2.11),

$$h(y, \theta) = h(y, p, \alpha, \beta, \gamma) = p_1 f_G(y, \gamma) + p_2 f_L(y, \alpha) + (1 - p_1 - p_2) f_W(y, \beta),$$

to the data from the 247 patients of Example 3.2. The same kind of priors and the same relationships among the model parameters (population mean and variance) were used, as well as a Dirichlet prior for the mixture parameters (p_1, p_2, p_3), with $p_1 + p_2 + p_3 = 1$. In this case, the p-values evaluated based on the Bayesian evidence indicate that neither the lognormal and gamma models should be considered because the null hypotheses $H : p_1 = 0$ and $H : p_2 = 0$ are not rejected (see Table 3.11). Consequently, among the three models, the Weibull model is the one that should be considered. From Table 3.11 and Fig. 3.5, it appears reasonable to disregard both the lognormal and gamma models; the Weibull model by itself produces a good estimate of the survival function (Tables 3.10 and 3.12).

3.4 Bibliographic Notes

A comparison of the alternative Bayes factors from a more theoretical and fundamental point of view has not been attempted in this book. For such discussions on the POBF, refer to Aitkin (1992, 1993), Aitkin et al. (2005), and Lindley (1993). The

Table 3.10 Estimates of the gamma-lognormal-Weibull (GLW) mixture model

Parameter	Estimate	SD	LB 95 %	UB 95 %
p_1-gamma	0.25	0.19	0.00	0.61
p_2-lognormal	0.30	0.20	0.00	0.68
p_3-Weibull	0.45	0.22	0.04	0.85
μ	12.81	0.98	11.14	14.80
σ^2	83.47	37.15	41.14	146.04

Table 3.11 Hypothesis testing for the mixture parameters of the GLW mixture model

Hypothesis	e-value	p-value
$p_1 = 0$	0.81	0.13
$p_2 = 0$	0.80	0.12
$p_3 = 0$	0.15	0.00

Table 3.12 Estimates of the Weibull model

Parameter	Estimate	SD	LB 95 %	UB 95 %
μ	12.40	0.69	11.15	13.82
σ^2	58.70	11.53	39.11	81.74

FBF and IBF were the focus of papers by O'Hagan (1997) and Berger and Mortera (1999) as well as a series of papers by Berger and Pericchi and by De Santis and Spezzaferri; all of these works are referenced in the review papers of Berger and Pericchi (2001) and Pericchi (2005). A general expression for deriving these Bayes factors is given by Gelfand and Dey (1994). Further simulation results on the FBF, IBF, and POBF were presented in the unpublished thesis of Araujo (1998). Another use of the Bayes factor is to order the sample space in any dimension and then use this order to define new standard p-values; see Pereira and Wechsler (1993), Pericchi and Pereira (2016).

Regarding the FBST, it was originally developed to test sharp hypotheses in both sample and parametric spaces of any dimensions. However, it can also be used for non-sharp hypotheses. We understand a sharp hypothesis to be a hypothesis that is defined in a subspace of a smaller dimensionality than the original parameter space. Madruga et al. (2001) proved the Bayesianity of the FBST and that, with suitable modification, the FBST becomes invariant under parametric transformations (see Madruga et al. 2003). This is not to be confused with the work of Box and Tiao (1965) on credible intervals, which only compared fixed credibility intervals with the hypothesis under study, looking for the intersection of the hypothesis with the credible region of a fixed credibility interval: there are an infinite number of hypotheses intersecting such regions. West and Harrison (1997, Sect. 17.3.5) and Basu (1996) also attempted to define such a test but only considered real-line spaces and did not correct for invariance under parametric transformations. The FBST is somewhat related to Barnard's OAAAA method, presented in Sect. 1.3. For papers that discuss the FBST and a probability value analogous to the frequentist concept of power, see Rogatko et al. (2002), Stern and Zacks (2002), Lauretto et al. (2007) and Isbicki et al. (2011). A paper demonstrating an additional functionality of this Bayesian test is that by Lauretto et al. (2003), in which the Behrens–Fisher problem is treated as a simple application of a general solution to many questions on multivariate normality.

References

Aitkin, M.: Posterior Bayes factors (with discussion). J. Roy. Stat. Soc. B **51**, 111–142 (1991)

Aitkin, M.: Evidence and the posterior factor. Math. Sci. **17**, 15–25 (1992)

Aitkin, M.: Posterior Bayes factor analysis for an exponential regression model. Stat. Comput. **3**, 17–22 (1993)

Aitkin, M., Boys, R.J., Chadwick, T.: Bayesian point null hypothesis testing via the posterior likelihood ratio. Stat. Comput. **15**, 215–230 (2005)

Araujo, M.I.: Comparison of separate models: a bayesian approach using improper prior distributions, Ph.D. Thesis (Operational Research). COPPE/UFRJ-Federal University of Rio de Janeiro (In Portuguese) (1998)

Araujo, M.I., Fernandes, M., Pereira, B.B.: Alternative procedures do discriminate nonnested multivariate linear regression models. Commun. Stat Theory Methods **34**, 2047–2062 (2005)

Araujo, M.I., Pereira, B.B.: Bayes factors to discriminate separate multivariate regression using improper prior (in Portuguese). Rev. Bras. Estatí. **68**, 33–50 (2007a)

Araujo, M.I., Pereira, B.B.: A comparison of Bayes factors for separated models: some simulation results. Commun. Stat. Simul. Comput. **36**, 297–309 (2007b)

Atkinson, A.C.: Posterior probabilities for choosing a regression model. Biometrika **65**, 39–48 (1978)

Barbosa, F.H.: The Brazilian inflaction in the post-war: monetarism versus structuralism. IPEA/INPES (1983) (In Portuguese)

Basu, S.A.: A new look at Bayesian point null hypothesis testing. Sankhya Indian J Stat. Ser. A **58**, 292–310 (1996)

Berger, J.O., Pericchi, L.R.: The intrinsic Bayes factor for model selection and prediction. J. Am. Stat. Ass. **91**, 109–122 (1996)

Berger, J.O., Mortera, J.: Default Bayes factor for nonnested hypotheses testing. J. Am. Stat. Ass. **94**, 542–554 (1999)

Berger, J., Pericchi, L.R.: Objective Bayesian model methods for model selection: introduction and comparison (with discussion). In: Lahir, P. (ed.) IMS Lecture Notes-Monograph Series, vol. 18, pp. 135–207 (2001)

Box, G.E.P., Tiao, G.C.: Multiparameter problems from a Bayesian point of view. Ann. Math. Stat. **36**, 1468–1482 (1965)

Cox, D.R.: Tests of separate families of hypotheses. In: Proceedings 4th Berkeley Symposium in Mathematical Statistics and Probability, vol. 1, pp. 105–123 (1961). University of California Press

Cox, D.R.: The role of significance tests. Scand. J. Stat. **4**, 49–70 (1977)

Diniz, M., Pereira, C.A.B., Polpo, A., Stern, J.M., Wechsler, S.: Relationship between Bayesian and frequentist significance indices. Int. J. Uncertainty Quantif. **2**, 161–172 (2012)

Gelfand, A.E., Dey, D.K.: Bayesian model choice: asymptotic and exact calculations. J. Roy. Stat. Soc. B **56**, 501–504 (1994)

Isbicki, R., Fossaluza, V., Hounie, A.G., Nakano, E.Y., Pereira, C.A.B.: Testing allele homogeneity: the problem of nested hypotheses. BMC Genet. **13**, 103 (2011)

Kass, R.E., Raftery, A.E.: Bayes factors. J. Am. Stat. Assoc. **90**, 773–795 (1995)

Kempthorne, O.: Of what use are tests of significance and tests of hypothesis. Commun. Stat. Theory Methods **8**, 763–777 (1976)
Lauretto, M.S., Pereira, C.A.B., Stern, J.M., Zacks, S.: Full Bayesian significance test applied to multivariate normal structure models. Braz. J. Probab. Stat. **17**, 147–168 (2003)
Lauretto, M.S., de Faria-Jr, S.P. Pereira, C.A.B., Pereira, B.B., Stern, J.: The problem of separate hypotheses via mixture models. In: Knuth, K.H., Caticha, A., Center, J.L., Jr, Griffin, A., Rodrigues, C.C.: Orgs, Bayesian and maximum entropy methods in science and engineering. In: American Institute of Physics Proceedings, vol. 954, pp. 268–275 (2007)
Lauretto, M., Nakano, F., Pereira, C.A.B., Stern, J.M.: A Straightforward multiallelic significance test for Hardy-Weinberg equilibrium law. Genet. Mol. Biol. **32**, 619–625 (2009)
Lempers, F.B.: Posterior probabilities of alternative linear models. University of Rotterdam Press (1971)
Lindley, D.V.: On the presentation of evidence. Math. Sci. **18**, 60–63 (1993)
Madruga, M.R., Esteves, L.G., Wechser, S.: On the Bayesianity of Pereira-Stern tests. Test **10**, 291–299 (2001)
Madruga, M.R., Pereira, C.A.B., Stern, J.M.: Bayesian evidence test for precise hypotheses. J. Stat. Plan. Inference **117**, 185–198 (2003)
Melo, B.A.R.: Bayesian analysis of finite mixture models with censure data and covariates. Ph.D, Qualification Report (in Portugues) (2016)
O'Hagan, A.: Fractional Bayes factor for model comparison (with discussion). J. Roy. Stat. Soc. B **1**, 99–138 (1995)
O'Hagan, A.: Properties of intrinsic and fractional Bayes factors. Test **6**, 101–108 (1997)
Pereira, C.A.B., Wechsler, S.: On the concept of P-Value. Braz. J. Probab. Stat. **7**, 159–177 (1993)
Pereira, C.A.B., Stern, J.M.: Evidence and credibility: full Bayesian significance test for precise hypothesis. Entropy **1**, 69–80 (1999)
Pereira, C.A.B., Stern, J.M., Wechsler, S.: Can a significance test be genuinely Bayesian. Bayesian Anal. **3**, 79–100 (2008)
Pereira, C.A.B., Polpo, A.: Bayesian statistics with applications to categorical and survival data (in portuguese- Estatística Bayesiana com Aplicações em Dados Categóricos e de Sobrevivência). RBRAS, Campina Grande, 61 p (2014)
Pericchi, L.R.: An alternative to standard Bayesian procedure for discrimination between normal linear models. Biometrika **71**, 575–581 (1984)
Pericchi, L.R., Pereira, C.A.B.: Adaptative significance level using optimal decision rule: Balancing by weighting the error probability. Braz. J. Probabil. Stat. **30**, 70–90 (2016)
Pericchi, L.R.: Model selection and hypothesis testing based on objective probabilities and Bayes factors. In: Dey, D.K., Rao, C.R., Handbook of Statistics: Bayesian Thinking. Modeling and Computation. North Holland, vol. 25, pp. 115–149 (2005)
Rogatko, A., Slifker, M.J., Babb, J.S.: Hardy-Weinberg equilibrium diagnostics. Theoret. Popul. Biol. **62**, 251–257 (2002)
Rust, R.T., Schmittlein, D.C.: A Bayesian cross-validated likelihood method for comparing alternative specifications of quantitative models. Mark. Sci. **4**, 20–40 (1985)
Smith, A.F.M., Spiegelhalter, D.J.: Bayes factors and choice criteria for linear models. J. Roy. Stat. Soc. B **42**, 213–220 (1980)
Spiegelhalter, D.J., Smith, A.F.M.: Bayes factors for linear and log-linear model with vague prior information. J. Roy. Stat. Soc. B **44**, 377–387 (1982)
Stern, J.M., Zacks, S.: Testing the independence of poisson variates under the holgate bivariate distribution: the power of new evidence teste. Stat. Probabil. Lett. **60**, 313–320 (2002)
West, M., Harrison, J.: Bayesian Forecasting and Dynamic Models, 2nd edn. Springer, New York (1997)

Chapter 4
Support and Simulation Methods

Abstract This chapter addresses the pure likelihood approach to model choice. The concepts of normalized, adjusted, relative, and profile likelihood are introduced. A relative likelihood approach for discriminating separate models is presented using an example. The concepts of computer simulations, the Monte Carlo method, Monte Carlo simulations, and bootstrapping are described. Linear and nonlinear regression models in the literature are used as illustrations. An example is presented to demonstrate the use of a likelihood dominance criterion (LDC) for model choice.

Keywords Adjusted likelihood · Bootstrap · Fisher approach · Generalized linear models · Histogram · Likelihood law · Likelihood dominance criterion · Neyman–Pearson approach · Normalized likelihood · Profile likelihood · Relative likelihood

4.1 Introduction

This chapter describes direct applications of the likelihood function as a measure of support for a model compared with an alternative separate model. The model with the greater support is the preferred model.

Finally, to overcome the difficulties encountered in obtaining analytical expressions for tests of separate hypotheses, this chapter also presents several simulation-based alternatives.

4.2 Likelihood Inference

The likelihood function plays a central role in parametric inference because it contains all information on the observed data. Although the likelihood figures prominently in all antagonistic Fisherian, Neyman–Pearson, and Bayesian views, it is not their main objective.

B. de B. Pereira and C.A. de B. Pereira,
Model Choice in Nonnested Families, SpringerBriefs in Statistics,
DOI 10.1007/978-3-662-53736-7_4

Bayesians and frequentists may disagree with the views presented here. These views are close to Fisher's ideas presented in his last and controversial book, Fisher (1956).

Several approaches to statistical tests exist (see Pereira and Pereira 2005): the Fisherian significance test, the Neyman–Pearson hypothesis test, and the FBST procedure (Sect. 3.3). Another approach to hypothesis testing is stated in terms of the Pure Likelihood Law. Edwards (1992) called it the Method of Support. The likelihood function, L, introduces an ordering of preferences of all possible parameter points. Note that this ordering remains the same when any proportional function is considered. This means that we can divide L by any constant, such as the integral or maximum of the likelihood function: the former is the Normalized Likelihood, the Bayesian way, and the latter is the Relative Likelihood (RL). These modified likelihoods are defined whenever the corresponding constants exist. This section ends by stating a rule to be used by Pure Likelihood followers:

"Pure Likelihood Law": If the relative likelihoods (RL) of hypotheses H_f and H_g satisfy $RL(H_f) > (<) RL(H_g)$, then we say that H_f is more (less) plausible than H_g. The strength of the evidence provided by the data y in favor of H_f against H_g is measured in terms of the likelihood ratio (LR).

$$LR(H_f, H_g) = RL(H_f)/RL(H_g). \tag{4.1}$$

Example 4.1 (Lindsey 1974a) Let independent observations $(y_1, \ldots, y_n)$ be summarized in a histogram with k bins with frequency n_j in bin j ($n = \sum_j^k n_j$) and theoretical proportion (probability) p_j. The best estimate of p_j is

$$\hat{p}_j = n_j / \sum_j^k n_j = n_j/n. \tag{4.2}$$

The probability of the observed data given the estimated proportions $\hat{p}_j$ is proportional to

$$L_M \hat{P} = \prod_j \hat{p}_j^{n_j}, \tag{4.3}$$

the likelihood of the multinomial model. For a proposed distribution for the data, the predicted proportion of observations falling into interval j, given the data, will be as follows:

- for discrete Y,

$$\tilde{p}_j = P(y_j, \hat{\theta}), \tag{4.4}$$

- for continuous Y,

$$\tilde{p}_j = \int_a^b f(y_j, \hat{\theta}) dy \approx f(y_j, \hat{\theta}) \Delta y_j, \tag{4.5}$$

where P and f are a probability and a density function, respectively; $\hat{\theta}$ is an estimate of the unknown parameter θ; $a = y_j - \frac{1}{2}\Delta y_j$; and $b = y_j + \frac{1}{2}\Delta y_j$.

The resulting likelihood functions are

$$L(\hat{\theta}) = \prod_{j=1}^{k} \tilde{p}_j^{n_j} = (\tilde{p}_j = \prod_{j,k} P(y_j, \hat{\theta}) \quad \text{discrete}, \tag{4.6}$$

$$L(\hat{\theta}) = \prod_{j=1}^{k} f(y_j, \hat{\theta}) \Delta y_j \quad \text{continuous}. \tag{4.7}$$

The plausibility or the support of the theoretical distributions (P *or* f) compared with the most plausible one is

$$RL = \prod_{j=1}^{k} (\tilde{p}_j / \hat{p}_j)^{n_j}. \tag{4.8}$$

For a Cox (1962) comparison of the Poisson and geometric distributions for 30 observations generated from a Poisson model, Lindsey (1974a) obtained

$$\begin{aligned} \tilde{p}_P &= \exp(-\hat{\theta})\hat{\theta}^y / y!, \\ \tilde{p}_G &= \hat{\theta}^y / (1 + \hat{\theta})^{1+y}. \end{aligned} \tag{4.9}$$

$\log RL_p = -0.609$ and $\log RL_G = -3.548$, which favor the Poisson model. In addition, Lindsey (1974b) presented an extension for regression models.

Example 4.2 (Pollack and Wales 1991) Consider a comprehensive model H_c that includes models H_f and H_g. Let k_j be the number of parameters in $H_j (j = f, g$ and $c)$.

Let ℓ_1, ℓ_2, and ℓ_c denote the log-likelihoods corresponding to the three hypotheses H_j, and let $C(v)$ be the value of a chi-squared distribution with v degrees of freedom at some fixed significance level.

Under the likelihood ratio test, the hypothesis H_i will not be rejected when tested against H_c if $2(\ell_c - \ell_i) < C(k_c - k_i)$.
Here, only the two outcomes

(a) reject H_f and accept H_g and
(b) reject H_g and accept H_f

of the four possible outcomes listed in Sect. 2.2.3 are considered. The following procedure eliminates the necessity of estimating or even specifying a particular comprehensive model if only outcomes (a) and (b) are of interest.

Suppose that the models specified by H_f and H_g are estimated and defined by the corresponding "adjusted likelihood values" $V_i = \ell_i + C(k_c - k_i)/2$. There are three possible cases.

First, suppose that $V_g > V_f$, and consider an imaginary experiment in which a particular comprehensive model with R_c parameters and its associated likelihood ℓ_c is estimated. The value of ℓ_c lies in one of three regions:

- if $\ell_c < V_f$, both H_f and H_g are accepted;
- if $\ell_f < \ell_c < \ell_g$, H_f is rejected and H_g is accepted;
- if $\ell_c > l_g$, both H_f and H_g are rejected.

Thus, if $V_g > V_f$, then there is no value of ℓ_c for which H_f is accepted and H_g is rejected.

Second, suppose that $V_f > V_g$. A similar argument shows that there is no value of ℓ_c for which H_g is accepted and H_f is rejected.

Finally, suppose that $V_f = V_g$. In this case, there are only two possibilities:

- if $2\ell_c$ is less than $V_f = V_g$, both H_f and H_g are accepted;
- if $2\ell_c$ is greater than $V_f = V_g$, both H_f and H_g are rejected in favor of H_c.

Thus, when $V_f = V_g$, there is no value of ℓ_c that would lead to accepting one hypothesis and rejecting the other.

Pollack and Wales (1991) called this procedure the "likelihood dominance criterion" (LDC) and suggested the following criteria (assuming that $k_f < k_g$ and $k_c = k_f + k_g + 1$):

(i) The LDC prefers H_f to H_g if $\ell_g - \ell_f < [C(k_f + 1) - C(k_g + 1)]/2$.
(ii) The LDC is indecisive between H_f and H_g if
$[C(k_g + 1) - C(k_f + 1)]/2 < \ell_f - \ell_g < [C(k_g - k_f + 1) - C(1)]/2$.
(iii) The LDC prefers H_g to H_f if $\ell_f - \ell_g > [C(k_g - k_f + 1) - C(1)]/2$.

The C values depend not only on k_f and k_g but also on the significance level chosen. The suggested value $k_c = k_f + k_g + 1$ arises from the exponential and linear combination of H_f and H_g from previous chapters.

Their paper ends with an application in the domain of consumer demand analysis, comparing the quadratic expenditure system and generalized translog models.

Example 4.3 (Cole 1975) Ventilatory function is a measure of the amount of air that an individual can breathe and is used for screening against chronic respiratory disease. Two indices used to quantify it are forced ventilatory volume (FEV) and force vital capacity (FVC), both derived from the volume of air in liters expired in a single forced expiration following a full inspiration. Both indices are larger in tall individuals and decline with age.

A reanalysis of nine studies of ventilatory function from all over the world involving more than 11000 men and women was conducted to select one of the following models:

$$\begin{aligned} 1:\ & FEV = a + b.age + c.height, \\ 2:\ & FEV = a + c.height + d.age.height, \\ 3:\ & FEV = c.height^m + d.age.height^m. \end{aligned} \tag{4.10}$$

If the parameter vector for model j is θ_j, then all models have the following form:

$$FEV = f(age, height, \theta_j) + \varepsilon_j,\ \ j = 1, 2, 3, \tag{4.11}$$

where the ε_j are assumed to follow $N(0, \sigma_j^2)$ distributions. For a sample of size n and apart from an arbitrary constant, the likelihood (or support, according to Edwards 1992 and as adopted by Cole 1975) is

$$S(\theta_j) = \ell(\hat{\theta}_j) = -\frac{n}{2} \ln \sigma_j^2 - \frac{1}{2\sigma_j^2} \sum \{FEV - E(FEV)\}^2. \tag{4.12}$$

Cole (1975) chose the appropriate model by comparing the values of $S(\theta_j)$ for all three models. A value of $m = 2$ was obtained by analyzing the profile likelihood:

$$S_3(m) = \arg \max_{c,d} S(c, d, m).$$

4.3 Simulations and Bootstrap

4.3.1 Simulations

Models are approximations of systems or processes and represent their key characteristics. Simulations emulate the operation of the system.

A mathematical model consists of algorithms and equations used to represent a structure that will reproduce the behavior of the system being modeled.

A computer simulation consists of the running of these equations and algorithms using high-speed computer power as a substitute for analytical calculations.

Monte Carlo methods or stochastic simulations are a class of computational algorithms that rely on random sampling to obtain numerical results. They are usually applied when it is impossible to obtain a closed-form expression or it is infeasible to apply a deterministic algorithm.

Sawilosky (2003) distinguishes between simulations, the Monte Carlo Method, and Monte Carlo simulations. A simulation is a fictional representation of reality, a numerical technique for conducting experiments. The Monte Carlo method is a stochastic technique for solving a deterministic problem, either a mathematical or a physical one. Monte Carlo simulations use repeated samples to determine the properties of a certain phenomenon.

4.3.2 Bootstrap

Bootstrapping refers to a metaphor for a self-sustained process that proceeds without external assistance. The term comes from the story "The Surprising Adventures of Baron Munchausen", in which the Baron pulls himself out of a swamp by his hair. The concept of bootstrapping arose from a variant of this tale.

In statistics , bootstrapping refers to one type of resampling method (others include Jackknife, Cross-validation, and Permutation Tests) that allows the estimation of the distribution of a statistic and measures the accuracy of the estimates. It is used to compute standard errors, confidence intervals, hypothesis and significance tests, and bias corrections.

It is especially useful, for example, when the statistic of interest is complicated, when the sample size of the study is small, or for the specification of a desired sample size based on pilot studies.

For observed data drawn from a random sample of size n, bootstrap yields a number B of resamples of the data set (with replacement), each with the same size n.

Inferences from the data (e.g., standard errors, confidence intervals, and statistical tests) can be obtained in the following ways:

(a) Nonparametric bootstrapping, based on the distribution $\hat{F}$: Instead of making inferences from the behavior of samples from F, the data-generating distribution, B samples are obtained from $\hat{F}$, the empirical distribution function.
(b) Parametric bootstrapping: A model is fitted to the data, usually using the maximum likelihood method, and B random samples are generated from this fitted model.

4.3.3 Applications

Example 4.4 (Williams 1970) Two regression models for the observed enzyme concentration y at time t are as follows:

(i) Segmented model f:

$$y_i = f(\alpha, t_i) + \varepsilon_{fi} \quad \alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, T_1, T_2), \tag{4.13}$$

where

$$\begin{aligned} f(\alpha, t) &= \alpha_0 + \alpha_1 t \quad for \;\; t \le T_1 \\ &= \alpha_0 + \alpha_1 t + \alpha_2 (t - T_1) \quad for \; T_1 \le t \le T_2 \\ &= \alpha_0 + \alpha_1 T_1 + \alpha_2 (T_2 - T_1) + \alpha_3 (t - T_2) \quad for \; T_2 \le t \end{aligned}$$

and the ε_{fi} are independently distributed with the form $N(0, \sigma_f^2)$.

(ii) Smooth model g:

$$y_i = g(\beta, t_i) + \varepsilon_{gi} \quad \beta = (\beta_0, \beta_1, \beta_2), \tag{4.14}$$

where $g(\beta, t_i) = \beta_0 + \beta_1 e^{\beta_3 t}$ and ε_{gi} are independently distributed with the form $N(0, \sigma_g^2)$.

The models were derived from two alternative theories regarding the synthesis of enzymes during the cell cycle.

Because one of the models presents discontinuities in its derivatives at unknown points T_1 and T_2, some difficulties arise in fitting and discriminating between alternative models.

Williams (1970) overcame these difficulties using a search procedure for the maximum likelihood estimation of the parameters of the segmented model and a simulation (parametric bootstrapping) to discriminate between the models. He used a discrimination criterion called the ratio of the maximized likelihood, λ:

$$\lambda = \frac{\text{residual sum of squares about the fitted segmented model}}{\text{residual sum of squares about the fitted smooth model}} \tag{4.15}$$

The likelihood of the segmented model is not differentiable with respect to all parameters; therefore, the Cox procedure (Sect. 2.2) cannot be applied. Instead, the simulation procedure described below was used.

Assuming that each model in turn is the true model, B samples of enzyme concentrations with size n are generated from $f(\hat{\alpha}, t_i) + \varepsilon_{fi}$ and $g(\hat{\beta}, t_i) + \varepsilon_{gi}$, where ε_{fi} and ε_{gi} are variates with $N(0, \hat{\sigma}_f^2)$ and $N(0, \hat{\sigma}_g^2)$ distributions, respectively. The variances $\hat{\sigma}_f^2$ and $\hat{\sigma}_g^2$ are obtained by dividing the residual sum of squares in the original sample by $n - 6$ and $n - 3$, respectively. Thus, B observations are drawn from each of two distributions Λ_f and Λ_g of λ_f and λ_g, respectively. The observation λ_0, namely, the value of λ obtained by fitting both regression models to the data, is to be allocated to one of the two distributions.

Let m_f, m_g, s_f, and s_g denote the means and standard deviations of the λ_{fi} and λ_{gi}, respectively. Let $d_f = max\{m_f + 2s_f, \max \lambda_{fi}\}$ and $d_g = min\{m_g - 2s_g, \min \lambda_{gi}\}$. Williams (1970) regarded λ_0 as a possible observation from Λ_f if $\lambda_0 < d_f$ and as a possible observation from Λ_g if $\lambda_0 > d_g$. Therefore, there are four possible conclusions:

- if $\lambda_0 < d_f$, $\lambda_0 < d_g$, the segmented model is chosen;
- if $\lambda_0 > d_f$, $\lambda_0 > d_g$, the smooth model is chosen;
- if $\lambda_0 > d_f$, $\lambda_0 < d_g$, both models are rejected; and
- if $\lambda_0 < d_f$, $\lambda_0 > d_g$, no discrimination between the two models is possible.

In one of his experiments, with $B = 10$, Williams (1970) ultimately obtained $\lambda_0 = 0.532$, $\hat{\sigma}_f = 5.33$ and $\hat{\sigma}_g = 7.14$. The 10 values of λ_{fi} and λ_{gi} were as follows:

λ_f: 0.549, 0.426, 0.437, 0.344, 0.508, 0.551, 0.461, 0.490, 0.423, 0.536;

λ_g: 1.213, 1.227, 1.269, 1.183, 1.264, 1.000, 0.998, 1.044, 0.951, 1.031.

The calculated value of λ was 0.532. Because this value lies within the range of λ_{fi} and well outside the range of λ_{gi}, the segmented model f was chosen.

Example 4.5 (Wahrendorf et al. 1987) Consider two models in the class of generalized linear models (see McCullagh and Nelder 1989): M_1, with r_1 parameters, and M_2, with r_2 parameters that are a subset of the parameters of model M_1 such that $r_2 < r_1$. Let $\ell(M_1)$ and $\ell(M_2)$ be the maximized likelihood functions of models M_1 and M_2, respectively. Under the null hypothesis H_0: the additional $r_1 - r_2$ parameters of model M_1 are all zero, and the likelihood ratio statistic is $LR(M_2, M_1) = -2\log\{\ell(M_2)/l(M_1)\} \sim \chi^2_{r_1-r_2}$, i.e., it follows a central chi-squared distribution with $r_1 - r_2$ degrees of freedom.

Under the alternative hypothesis H_1: at least one of the additional parameters is nonzero, $LR(M_2, M_1) = \chi(\delta)$, i.e., it follows a noncentral chi-squared distribution with a noncentrality parameter δ. Therefore, the null hypothesis can also be expressed as $\delta = 0$.

Consider two models f and g with r parameters in common and r_f and r_g additional parameters, respectively. The model with only the r common parameters is denoted by M_{fg}.

If not all $r_f(r_g)$ parameters of models $f(g)$ are zero, $L(M_{fg}, f) \sim \chi^2(\delta_f)$ (similarly $L(M_{fg}, g) \sim \chi^2(\delta_g)$).

Consider the case in which $r_f = r_g$. The improvements in fit (over model M_{fg}) offered by model f or model g can be compared by testing the hypothesis $\delta_f = \delta_g$ against $\delta_f \neq \delta_g$. If $r_f \neq r_g$, then the interpretation of the difference in the noncentrality parameters is ambiguous.

Wahrendorf et al. (1987) showed that $\hat{\delta}_f = \hat{\delta}_g$ if and only if $LR(f, F) = LR(g, F)$, where F is a full model with all $r + r_f + r_g$ parameters. These likelihood ratios are the deviances of a generalized linear model.

Note that $LR(f, F)$ and $LR(g, F)$ are not independent. Thus, to test whether two nonnested models with equal degrees of freedom fit the data equally well is equivalent to testing $\delta_f = \delta_g$ or $LR(f, F) = LR(g, F)$. The null distribution of these statistics is the distribution of the difference of two dependent χ^2 distributions with equal degrees of freedom.

To perform the test above, we need to calculate the sample distribution of the test under the null hypothesis. Because this would be difficult to accomplish analytically, the authors used the bootstrap technique to estimate the sample distribution of the difference of the deviances given the observations.

This author applied the above procedure to two sets of data as follows:

(a) The nonparametric bootstrap approach was used in carcinogenesis dose-response experiments to choose among alternative Cox regression models (Cox 1972) that were fitted to the survival times of groups of mice treated with different doses of an initiator and a promoter used in a standard fashion. The times of occurrences of papillomas were monitored and used as endpoints in a censored failure time analysis. The hazard function was $\lambda(t) = \lambda_0 e^{\theta z}$, and the models were $f : z_f = (dose, \sqrt{dose})$, $g : z_g = (dose, \log dose)$, and $M_{fg} : z_M = (dose)$. Upon performing a bootstrap experiment with $B = 1000$, the histogram and confidence intervals, to perform hypothesis and significance tests, and to carry out bias corrections, for the differences of the deviances indicated that model g was not better than model f.

(b) A parametric bootstrap approach to choose between additive or multiplicative models was used on data regarding deaths from coronary heart disease among British male doctors. The number of deaths was considered to be a Poisson random variate. For the division of the data according to 5 age categories and the presence or absence of a smoking habit, the models were as follows for the death rates λ_{jk} in age groups j $(j = 1, ..., 5)$ for nonsmokers $(k = 0)$ and smokers $(k = 1)$, with covariates $z = 0$ and $z = 1$, respectively:

(i) Multiplicative:

$$f : \lambda_{jk} = \lambda_{j0} \exp(\alpha\beta) = \exp(\alpha_j + \alpha_z), \tag{4.16}$$

where λ_{j0} is an age-specific rate λ_{j0}.

(ii) Additive:

$$g : \lambda_{jk} = \beta_j + \beta_z. \tag{4.17}$$

Bootstrap samples were generated using the observed values as parameters of independent Poisson distributions. The multiplicative and additive models were fitted to each bootstrap sample, and their respective deviances were computed. For the $B = 1000$ bootstrap samples, the distribution of the differences of the deviances between the multiplicative and additive models was found to be symmetric, with a mean of 9.97 and standard deviation of 7.75. The bootstrap confidence intervals for the deviance differences may be attributed to chance; that is, model g is not better than model f.

Example 4.6 (Schork and Schork 1989, Example 1.3 cont.) The basic bootstrap method described by the authors is the same as that presented in Example 4.1. Here, $H_f : f(y, \alpha)$ is a lognormal density, and $H_g : g(y, \beta)$ is a mixture of two normal densities. For a sample $(y_1, ..., y_n)$,

$$\hat{\lambda} = \sum_{i=1}^{n} \log g(y_i, \hat{\beta}/f(y_i, \hat{\alpha}). \tag{4.18}$$

The following procedure illustrates the basic motivation behind a parametric bootstrap test. To test H_f, generate B samples of size n from the density $f(y, \hat{\alpha})$, and for each sample, estimate $\hat{\alpha}*$, $\hat{\beta}*$, and $\hat{\lambda}^*$. Critical values for the nonartificial $\hat{\lambda}$ can be obtained from the order statistics of artificial $\hat{\lambda}^*$s.

This procedure was used by (Schork and Schork 1989) to test alternative genetic hypotheses. The Pickering/Plat debate described in Example 1.1 is discussed in Schork et al. (1990). The data consisted of systolic and diastolic blood pressure values collected from 941 white male subjects participating in a random blood pressure trial at Michigan State University. The data were adjusted for the effects of age, height, and weight. The differences in the log-likelihoods were 23.53 for systolic pressure and 4.46 for diastolic pressure. The critical values of the parametric bootstrap test at a 5% level of significance were 2.75 and 3.02. Therefore, the lognormal distribution was rejected, and there found to be a greater potential for a normal mixture, corresponding to the genetic hypothesis for hypertension. Note that this analysis of the blood distribution is not intended as an exhaustive resolution to the issues raised in the Pickering/Plat debate.

Example 4.7 (Nevill and Holder 1994) Maximum oxygen uptake ($VO_2(max)$) is a measure of an individual's capacity to deliver oxygen to and use oxygen in an exercised muscle. It is considered an important single indicator of cardiovascular fitness. It is known that several factors affect $VO_2(max)$, such as body size, age, gender, and the amount of exercise that the individual performs.

Analyzing data on 1732 subjects from the Allied Dunbar National Fitness Survey (ADNFS), (Nevill and Holder 1994) adapted and generalized the FEV model of Cole (1975) presented in Example 4.3 as follows:

$$\begin{aligned} FEV &= height^k(c + d.age) + \varepsilon, \\ VO_2(max) &= weight^k(c + d.age) + \varepsilon. \end{aligned} \tag{4.19}$$

They also incorporated dichotomous variables of gender (z) and vigorous exercise (v) by allowing parameters k and c to vary. The model was thus as follows:

$$\begin{aligned} VO_2(max) = weight^{k_0+k_1z+k_2v+k_3zv}\{&c_0 + c_1z + c_2v + \\ &c_3zv + (d_0 + d_1z + d_2v + d_3zv)age\} + \varepsilon. \end{aligned} \tag{4.20}$$

The authors also considered the following multiplicative model, which they believed to be more plausible:

$$VO_2(max) = weight^k \exp(c + d.age)\varepsilon. \tag{4.21}$$

Incorporating gender and vigorous exercise results into the log-linear model yielded

$$\begin{aligned} \log VO_2(max) = (k_0 + k_1z + k_2v + k_3zv) \log weight + c_0 + c_1z + c_2v + \\ c_3zv + (d_0 + d_1z + d_2v + d_3zv)age + \varepsilon. \end{aligned} \tag{4.22}$$

Because of the noted heteroscedasticity of the data, the authors proceeded to estimate the models by assuming normality for the error terms, using weighted regression, and minimizing

$$\begin{aligned} &\tfrac{1}{n}\sum w_i(y_i - f(x,\alpha))^2, \\ &\tfrac{1}{n}\sum w_i(\ln y_i - g(x,\beta))^2, \end{aligned} \tag{4.23}$$

for the nonlinear and log-linear models, respectively.

Finally, using the bootstrap approach of the previous examples, they chose the log-linear model. It was also noted that the residuals from the nonlinear model deviated considerably from normality.

Example 4.8 (Cribari-Neto and Lucena 2015, Example 2.11 cont.)

The authors performed bootstrap versions of the likelihood ratio and Wald tests to test the five models in (2.48). Their procedure was as follows:

(i) Estimate all models m_i ($m_i \neq m_f$), obtain the $\hat{\eta}_i$ ($i \neq f$), include them as additional covariates in model m_f, and estimate the resulting augmented model.
(ii) Compute the J statistic.
(iii) Generate a bootstrap sample of the response y_f^* from model m_f.
(iv) Estimate the augmented model using y_f^* as the response and compute J^*.
(v) Repeat (iii) and (iv) B times, where B is a large positive integer.
(vi) Compute $T_{1-\alpha}$, the $1-\alpha$ quantile of the B bootstrap statistics (J_1^*, ..., J_B^*).
(vii) Reject m_f if $J > T_{1-\alpha}$.

For testing $m_j \neq m_i$, proceed similarly.

For the bootstrap MJ statistic, they proceeded as follows:

(i) Calculate the MJ statistic as described above.
(ii) Generate a bootstrap sample of the response y_i^* as above.
(iii) Calculate the MJ bootstrap statistics, MJ^*.
(iv) Repeat (ii) and (iii) B times.
(v) Compute $T_{1-\alpha}$, the $1-\alpha$ quantile of the B bootstrap statistics (MJ_1^*, ..., MJ_B^*).
(vi) Reject the null hypothesis that the true model belongs to the set of candidate models if $MJ > T_{1-\alpha}$.

The decision rule can also be expressed in terms of the bootstrap p-value, which is given by p^*, the proportion of times that the bootstrap statistic, say MJ_b^* ($b = 1, ..., B$), is smaller than the selected nominal level.

The J and MJ tests were performed by the authors for both the likelihood ratio and Wald tests and their bootstrap versions. The p-values of the J tests for pairwise nonnested models are reported in Table 4.1.

The MJ bootstrap values were 0.4366 and 0.2867 for the likelihood ratio and Wald bootstraps, respectively. Therefore, the correct model was concluded to be among the candidate models, and because the smallest J statistic was that of the log–log model, this model was selected based on the MJ test.

Table 4.1 p-values for the J test obtained using the likelihood ratio (LR) and Wald statistics for the five competing models; the bootstrap p-values are also reported

Model	LR	LR_{boot}	Wald	$Wald_{boot}$
Logit versus probit	1.715×10^{-5}	0.0060	2.637×10^{-8}	0.0050
Logit versus log–log	1.828×10^{-5}	0.0040	2.657×10^{-8}	0.0110
Logit versus compl. log–log	0.0004	0.0190	1.667×10^{-5}	0.0025
Logit versus Cauchit	0.0023	0.06190	0.0003	0.0639
Probit versus logit	0.0016	0.0150	0.0007	0.0140
Probit versus log–log	0.0040	0.0070	0.0001	0.0040
Probit versus compl. log–log	0.0026	0.0190	0.0013	0.0160
Probit versus Cauchit	0.0089	0.0470	0.0061	0.0499
Log–log versus logit	0.4869	0.6074	0.4863	0.5614
Log–log versus probit	0.2634	0.3646	0.2596	0.3926
Log–log versus compl. log–log	0.5505	0.6414	0.5501	0.6234
Log–log versus Cauchit	0.7583	0.8092	0.7584	0.8232
Compl. log–log versus logit	1.629×10^{-5}	0.0060	8.207×10^{-9}	0.0090
Compl. log–log versus probit	8.581×10^{-7}	0.0030	3.25×10^{-12}	0.0010
Compl. log–log versus log–log	1.496×10^{-6}	0.0010	9.013×10^{-12}	0.0010
Compl. log–log versus Cauchit	0.0030	0.0460	6.319×10^{-6}	0.0260
Cauchit versus logit	5.4×10^{-8}	0.0080	2.028×10^{-12}	0.0010
Cauchit versus probit	6.01×10^{-9}	0.0020	2.527×10^{-15}	0.0010
Cauchit versus log–log	1.6×10^{-10}	0.0200	$<2.2 \times 10^{-16}$	0.0010
Cauchit versus compl. log–log	2.193×10^{-7}	0.0240	6.624×10^{-11}	0.0010

4.4 Bibliographic Notes

There has been a recent revival of interest in the likelihood inference method of Fisher (1956). Readable accounts can be found in Edwards (1992), Kalbfleisch (2011), King (1998), Lindsey (1986), Pawitan (2001), Royall (1999), and Sprott (2000). For the application of this method in clinical medicine, refer to Pereira and Pereira (2005), and time series in Barnard et al. (1962).

The reasoning underlying the support method also serves as the foundation for the OAAAA method of Barnard, introduced in Sect. 1.3, and the FBST, introduced in Sect. 3.3. In an unpublished thesis, Rojas (2001) used simulations to determine the probability of correct selection using the support method for the exponential, Weibull, gamma, and lognormal distributions.

In Jackson (1967, 1968), Pereira (1976, 1981) and Loh (1985), simulation results were used to study the significance levels of accuracy, power properties, and convergence to normality of statistical procedures for testing separate hypotheses.

More recently, simulations have been used as computer-intensive methods of testing these hypotheses, and the examples presented in this chapter are representative of such efforts.

Results in bootstrap theory suggest that the use of asymptotically pivotal quantities (APQs), namely, random variables whose asymptotic distributions do not depend on any parameters, leads to procedures with a higher level of accuracy. The Cox statistic introduced in Chap. 2 is an APQ.

Schork (1993), Pesaran and Pesaran (1993, 1995), and Coulibaly and Brorsen (1999) presented alternative bootstrap APQs as approximations to the Cox statistic when it is difficult to obtain the expression for the Cox test.

These authors used a descriptive statistic (or nonparametric estimate) of the log-likelihood difference and its expectation under hypotheses H_f and H_g. Usually, a two-step procedure is applied to estimate the probability limit of the alternative model under the null model, that is, β_α and α_β under H_f and H_g, respectively.

Such simulations should be constrained to values near the original sample estimates $\hat{\alpha}$ (or $\hat{\beta}$) and $\beta_{\hat{\alpha}}$ (or $\alpha_{\hat{\beta}}$), as also suggested by Cox (2013).

Further references on the use of bootstrapping for J-type tests of separate hypotheses are presented in Cribari-Neto and Lucena (2015).

References

Barnard, G.A., Jenkins, G.M., Winstein, C.B.: Likelihood inference and time series (with discussion). J. Roy. Stat. Soc. B **13**, 351–352 (1962)

Cole, T.J.: Linear and proportional regression models in the prediction of ventilatory function (with discussion). J. Roy. Stat. Soc. A **138**, 297–338 (1975)

Coulibaly, N., Brorsen, B.W.: A Monte Carlo sampling approach to testing nonnested hypotheses: Monte Carlo results. Econ. Rev. **18**, 195–209 (1999)

Cox, D.R.: Further results on test of separate families of hypotheses. J. Roy. Stat. Soc. **B**, 406–424 (1962)

Cox, D.R.: Regression models and life table (with discussion). J. Roy. Stat. Soc. B **34**, 187–200 (1972)

Cox, D.R.: A return to an old paper: tests of separate families of hypotheses (with discussion). J. Roy. Stat. Soc. B **75**, 207–215 (2013)

Cribari-Neto, F., Lucena, S.E.F.: Nonnested hypothesis testing in the class of varying disperson beta regressions. J. Appl. Stat. **42**, 967–985 (2015)

Edwards, A.W.F.: Likelihood. The John Hopkins University Press (1992)

Fisher, R. A.: Statistical Methods and Scientific Inference. Hafner Press Macmillan Publishing (1956)

Jackson, O.A.Y.: Some Tests Associated with the Exponential Distribution. Ph.D Thesis (Statistics), Imperial College, University of London (1967)

Jackson, O.A.Y.: Some results on tests of separate families of hypotheses. Biometrika **55**, 355–363 (1968)

Kalbfleisch, J.G.: Probability and Statistical Inference, vol. 2. Springer, Satatistical Inference (2011)

King, G.: Unifying Political Methodology: The Likelihood Theory of Statistical Inference. University of Michigan Press (1998)

Lindsey, J.K.: Comparison of probability distributions. J. Roy. Stat. Soc. B **36**, 38–47 (1974a)

Lindsey, J.K.: Construction and comparison of statistical models. J. Roy. Stat. Soc. B **36**, 418–425 (1974b)

Lindsey, J.K.: Parametric Statistical Inference. Oxford University Press (1986)

Loh, W.Y.: A new method for testing separate families of hypotheses. J. Am. Stat. Assoc. **80**, 362–368 (1985)

McCullag, P., Nelder, J.: Generalized Linear Models. 2nd ed. Chapman and Hall (1989)

Nevill, A.M., Holder, R.L.: Modelling maximum oxygen uptake-a case study in non-linear regression model formulation and comparison. J. Roy. Stat. Soc. C (Appl. Stat.) **43**, 653–666 (1994)

Pawitan, Y.: In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford University Press (2001)

Pereira, B.B.: Some Results on Test of Separate Families of Hypotheses. Ph.D Thesis (Statistics), Imperial College, University of London (1976)

Pereira, B.B.: Empirical comparisons of some tests of separate families of hypotheses. Metrika **25**, 219–234 (1981)

Pereira, B.B., Pereira, C.A.B.: A likelihood approach to diagnostic test in clinical medicine. REVSTAT-Stat. J. **3**, 77–98 (2005)

Pesaran, M.H., Pesaran, B.: A simulation approach to the problem of computing Cox's statistics for testing nonested models. J. Econ. **57**, 377–392 (1993)

Pesaran, M.H., Pesaran, B.: A nonnested test of level differenced versus log-differenced stationary models. Econ. Rev. **14**, 213–227 (1995)

Pollack, R.A., Wales, T.J.: The likelihood dominance criterion: a new method to model selection. J. Econ. **47**, 227–242 (1991)

Rojas, F.A.R.: Evaluation of test of separate hypotheses, Ph.D. Thesis (Operational Research). COPPE/UFRJ-Federal University of Rio de Janeiro (2001) (In Portuguese)

Royall, R.M.: Statistical Evidence: A likelihood Paradigm. Chapman & Hall (1999)

Sawilosky, S.S.: You think you've got trivials. J. Modern Appl. Stat. Methods **2**, 218–225 (2003)

Schork, N.: Combining Monte Carlo and Cox tests of non-nested hypotheses. Commun. Stat. Simul. **24**, 939–954 (1993)

Schork, N., Schork, M.A.: Testing separate families of segregation hypotheses: Bootstrap methods. Am. J. Human Genet. **45**, 803–813 (1989)

Schork, N., Weder, A.B., Schork, M.A.: On the asymmetry of biological frequency distributions. Genet. Epidemiol. **7**, 427–446 (1990)

Sprott, D.A.: Statistical Inference in Science. Springer, New York (2000)

Wahrendorf, J., Becher, H., Brown, C.C.: Bootstrap comparison of non-nested generalized linear models: applications in survival anylis and epidemiology. Appl. Stat. **36**, 72–81 (1987)

Williams, D.A.: Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell culture. Biometrics **70**, 23–32 (1970)

Appendix A
Maximum Likelihood Estimation (MLE)

The results of the maximum likelihood estimation (MLE) of the lognormal, Weibull, gamma and exponential distributions, and regression models are presented; specifically, these results include the log-likelihood functions, the estimation equations, and the Fisher's information matrices. The notation is the same as that used in Examples 2.1, 2.2, and 2.3.

A.1 Lognormal Models

(i) Distribution
The corresponding density function is denoted by $f_L(y;\alpha_1,\alpha_2)$.

$$\ell_L(\alpha_1,\alpha_2;y) = -\frac{n}{2}\log\alpha_2 - n\log\sqrt{2\pi} - \sum_{i=1}^{n}\log y_i - \frac{1}{2\alpha_2}\sum_{i=1}^{n}(\log y_i - \alpha_1)^2,$$

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n}\log y_i}{n}, \quad \hat{\alpha}_2 = \frac{\sum_{i=1}^{n}(\log y_i - \hat{\alpha}_i)^2}{n}, \tag{A.1}$$

$$I(\alpha_1,\alpha_2) = n\begin{bmatrix} 1/\alpha_2 & 0 \\ 0 & 1/(2\alpha_2^2) \end{bmatrix}.$$

(ii) Regression
The corresponding density function is denoted by $f_L(y_i;\alpha_1,\alpha_2,\underset{\sim}{a}')$.

B. de B. Pereira and C.A. de B. Pereira,
Model Choice in Nonnested Families, SpringerBriefs in Statistics,
DOI 10.1007/978-3-662-53736-7

$$\ell_L(\alpha_1, \alpha_2, \underset{\sim}{a}'; \underset{\sim}{y}) = -\frac{n}{2}\log\alpha_2 - n\log\sqrt{2\pi} - \sum_{i=1}^{n}\log y_i$$
$$-\frac{1}{2\alpha_2}\sum_{i=1}^{n}(\log y_i - \alpha_1 - \underset{\sim}{z_i}\underset{\sim}{a})^2,$$

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n}\log y_i}{n}, \quad \hat{\underset{\sim}{a}} = (Z'Z)^{-1}\underset{\sim}{Z}L, \hat{\alpha}_2 = \frac{1}{n}(\underset{\sim}{L} - \alpha_1\underset{\sim}{l} - \underset{\sim}{Z}\hat{\underset{\sim}{a}})^1(\underset{\sim}{L} - \underset{\sim}{\hat{\alpha}}_1 l - \underset{\sim}{Z}\hat{\underset{\sim}{a}}),$$

$$I(\alpha_1, \alpha_2, \underset{\sim}{a}') = \begin{bmatrix} I(\alpha_1, \alpha_2) & 0 \\ 0 & \frac{1}{\alpha_2}Z'Z \end{bmatrix}. \tag{A.2}$$

A.2 Weibull Models

(i) Distribution
The corresponding density function is denoted by $f_W(y; \beta_1, \beta_2)$.

$$\ell_W(\beta_1, \beta_2; \underset{\sim}{y}) = n\log\beta_2 - n\beta_2\log\beta_1 + (\beta_2 - 1)\sum_{i=1}^{n}\log y_i - \sum_{i=1}^{n}\left(\frac{y_i}{\beta_1}\right)^{\beta_2},$$

$$\hat{\beta}_1^{\hat{\beta}_2} = \frac{\sum_{i=1}^{n} y_i^{\hat{\beta}_1}}{n}, \quad \hat{\beta}_2 = \left[\frac{\sum_{i=1}^{n} y_i^{\hat{\beta}_2}\log y_i}{\sum_{i=1}^{n} y_i^{\hat{\beta}_2}} - \frac{\sum_{i=1}^{n}\log y_i}{n}\right]^{-1}, \tag{A.3}$$

$$I(\beta_1, \beta_2) = \begin{bmatrix} \left(\frac{\beta_2^2}{\beta_1}\right) & -\frac{\psi(2)}{\beta} \\ -\frac{\psi(2)}{\beta_1} & \frac{\psi'(1)+\{\psi(2)\}^2}{\beta_2^2} \end{bmatrix}.$$

(ii) Regression
The corresponding density function is denoted by $f_W(y_i; \beta_1, \beta_2, \underset{\sim}{b}')$.

$$\ell_W(\beta_1, \beta_2, \underset{\sim}{b}'; \underset{\sim}{y}) = n\log\beta_2 - n\beta_1\beta_2 + (\beta_2 - 1)\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\left(\frac{y_i}{e^{\beta_1 + z_i\underset{\sim}{b}}}\right)^{\beta_2},$$

$$\sum_{i=1}^{n}\underset{\sim}{z_i}'\left(\frac{y_i}{e^{\underset{\sim}{z_i}\hat{\underset{\sim}{b}}}}\right)^{\hat{\beta}_2} = 0, \quad \hat{\beta}_2^{-1} = \frac{\sum_{i=1}^{n}\left(\frac{y_i}{e^{\underset{\sim}{z_i}\hat{\underset{\sim}{b}}}}\right)^{\beta_2}\log y_i}{\sum_{i=1}^{n}\left(\frac{y_i}{e^{\underset{\sim}{z_i}\hat{\underset{\sim}{b}}}}\right)^{\beta_2}} - \frac{\sum_{i=1}^{n}\log y_i}{n}, \tag{A.4}$$

$$\sum_{i=1}^{n}\left(\frac{y_i}{e^{\underset{\sim}{z_i}\hat{\underset{\sim}{b}}}}\right)^{\hat{\beta}_2} - n e^{\hat{\beta}_1\hat{\beta}_2} = 0,$$

$$I(\beta_1, \beta_2, \underset{\sim}{b}') = \begin{bmatrix} n\beta_2^2 & -n\psi(2) & 0 \\ -n\psi(2) & n\frac{\psi'(1)+\{\psi(2)\}^2}{\beta_2^2} & 0 \\ 0 & 0 & \beta_2^2 Z'Z \end{bmatrix}.$$

A.3 Gamma Models

(i) Distribution
The corresponding density function is denoted by $f_G(y_i; \gamma_1, \gamma_2)$.

$$\ell_G(\gamma_1, \gamma_2; \underset{\sim}{y}) = -n\log\Gamma(\gamma_2) + n\gamma_2\log\frac{\gamma_2}{\gamma_1} + (\gamma_2 - 1)\sum_{i=1}^{n}\log y_i - \frac{\gamma_2}{\gamma_1}\sum_{i=1}^{n} y_i,$$

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^{n} y_i}{n}, \quad \log\hat{\gamma}_2 - \psi(\hat{\gamma}_2) = \log\hat{\gamma}_1 - \frac{\sum_{i=1}^{n}\log y_i}{n}, \tag{A.5}$$

$$I(\gamma_1, \gamma_2) = \begin{bmatrix} \frac{\gamma_2}{\gamma_1^2} & 0 \\ 0 & \psi(\gamma_2) - \frac{1}{\gamma_2} \end{bmatrix}.$$

(ii) Regression
The corresponding density function is denoted by $f_G(y_i; \gamma_1, \gamma_2, \underset{\sim}{g})$.

$$\ell_G(\gamma_1, \gamma_2, \underset{\sim}{g}'; \underset{\sim}{y}) = -n\log\Gamma(\gamma_2) + n\gamma_2\log\gamma_2 - n\gamma_1\gamma_2 + (\gamma_2 - 1)\sum_{i=1}^{n}\log y_i$$
$$-\gamma_2\sum_{i=1}^{n}\frac{y_i}{e^{\gamma_1 + \underset{\sim}{z_i}\underset{\sim}{g}}},$$

$$\sum_{i=1}^{n}\frac{y_i}{e^{\underset{\sim}{z_i}\hat{\underset{\sim}{g}}}} - n e^{\hat{\gamma}_1} = 0, \quad \sum_{i=1}^{n}\underset{\sim}{z_i}'\frac{y_i}{e^{\underset{\sim}{z_i}\hat{\underset{\sim}{g}}}} = \underset{\sim}{0}', \tag{A.6}$$

$$\log\hat{\gamma}_2 - \psi(\hat{\gamma}_2) = \hat{\gamma}_1 - \frac{\sum_{i=1}^{n}\log y_i}{n},$$

$$I(\gamma_1, \gamma_2, \underset{\sim}{g}') = \begin{bmatrix} n\gamma_2 & 0 & \underset{\sim}{0} \\ 0 & n\left\{\psi(\gamma_2) - \frac{1}{\gamma_2}\right\} & \\ & \underset{\sim}{0}' & \gamma_2 Z'Z \end{bmatrix}.$$

A.4 Exponential Models

Exponential models are special cases of Weibull ($\beta_2 = 1$) and gamma ($\gamma_2 = 1$) models; therefore, the corresponding results can be obtained from the results for either of these.

(i) Distribution
The corresponding density function is denoted by $f_E(y_i; \delta)$.

$$\ell_E(\delta, \underset{\sim}{y}) = -n \log \delta - \frac{1}{\delta} \sum_{i=1}^{n} y_i,$$

$$\hat{\delta} = \frac{\sum_{i=1}^{n} y_i}{n}, \tag{A.7}$$

$$I(\delta) = \frac{n}{\delta^2}.$$

(ii) Regression
The corresponding density function is denoted by $f_E(y, \delta, \underset{\sim}{d}')$.

$$\ell_E(\delta, \underset{\sim}{d}'; \underset{\sim}{y}) = -n\delta - \sum_{i=1}^{n} \frac{y_i}{e^{\delta + \underset{\sim}{z_i} \underset{\sim}{d}}},$$

$$\sum_{i=1}^{n} \frac{y_i}{e^{\underset{\sim}{z_i} \hat{\underset{\sim}{d}}}} - n e^{\hat{\delta}} = 0, \quad \sum_{i=1}^{n} \underset{\sim}{z_i}' \frac{y_i}{e^{\underset{\sim}{z_i} \hat{\underset{\sim}{d}}}} = \underset{\sim}{0}, \tag{A.8}$$

$$I(\delta, \underset{\sim}{d}') = \begin{bmatrix} n & \underset{\sim}{0} \\ \underset{\sim}{0}' & Z'Z \end{bmatrix}.$$

A.5 Location-Scale Models

Finally, there is a further property of the maximum likelihood estimator that is also used frequently, which is useful for identifying the crucial parameters for tests based on the maximum likelihood ratio and, consequently, for determining the parameters to be varied in simulation studies.

The previously discussed models can also be written in the forms presented below:

$$\frac{1}{\sigma} f\left(\frac{x - \alpha}{\sigma}; q\right) \quad \text{or} \quad f(x - \alpha; \sigma, q). \tag{A.9}$$

It can be shown that for models of these forms, the distribution of the maximum likelihood ratio depends only on q or (σ, q), respectively. If the models are in the location-scale form $\frac{1}{\sigma} f(\frac{x-\alpha}{\sigma})$, then the maximum likelihood ratio distribution is independent of the parameters (Antle and Bain 1969).

Reference

Antle, C.E., Bain, L.J.: A property of maximum likelihood estimators of location and scale parameters. SIAM Review **11**, 251–253 (1969)

Index

B. de B. Pereira and C.A. de B. Pereira,
Model Choice in Nonnested Families, SpringerBriefs in Statistics,
DOI 10.1007/978-3-662-53736-7