Annals of the Institute of Statistical Mathematics (2020) 72:329–352
https://doi.org/10.1007/s10463-018-0691-8

Strong model dependence in statistical analysis: goodness of fit is not enough for model choice

John Copas[1] · **Shinto Eguchi**[2]

Received: 2 February 2018 / Revised: 25 June 2018 / Published online: 3 October 2018

Abstract
Most statistical methods are based on models, but most practical applications ignore the fact that the results depend on the model as well as on the data. This paper examines the size of this model dependence, and finds that there can be very considerable variation between the results of fitting different models to the same data, even if the models being considered are restricted to those which give an acceptable fit to the data. Under reasonable regularity conditions, we show that different empirically acceptable models can give rise to non-overlapping confidence intervals for the same parameter. Application papers need to recognize that the validity of conventional statistical results rests on the assumption that the underlying model is known to be correct, and that this is a much stronger requirement than merely confirming that the model gives a good fit to the data. The problem of model dependence is only partially resolved by using formal methods of model selection or model averaging.

Keywords Goodness-of-fit · Model choice · Model uncertainty · Subset selection

1 Introduction

Most statistical methods are based on a model, but the reasons for choosing any particular model are often less than convincing. As Hodges (1987) points out, models usually used in statistical practice are '... little more than conventions: they have become conventional through constant exposition in service courses and textbooks,

The online version of this article contains supplementary material.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-018-0691-8) contains supplementary material, which is available to authorized users.

✉ John Copas
jbc@stats.warwick.ac.uk

[1] Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

[2] Institute of Statistical Mathematics, Midori-cho 10-3, Tachikawa, Tokyo 190-8562, Japan

through availability in popular software packages and because their mathematical tractability makes them inviting examples ...'. However, if we plot some relevant aspects of the data and find good agreement with what we would expect under a fitted model, then we usually take this as confirmation that the model, and hence our inference based on it, is at least reasonable. The tacit assumption is that if a model gives an acceptable fit to the data then inferences based on it will be sound. Also, we like our models to be simple because they tend to give narrower confidence intervals and less uncertainty. This amounts to a second tacit assumption that a simple model is to be preferred unless the data indicate a clear need for greater complexity. It is the consequences of these tacit assumptions that we want to discuss.

For any given dataset, there will usually be many models M that give an acceptable fit to the data and so would appear to be reasonable models for the analysis. Different Ms give different inferences. Imagine a paper reporting a clinical trial showing a significant treatment effect for some new drug. The author's analysis is based on model M_1. The paper is submitted to a medical journal and sent to two referees. Referee 1 checks the author's analysis and confirms that it is correct. Referee 2 decides to do his own independent re-analysis of the data and disagrees with the author—this referee uses model M_2 and finds that the treatment effect is no longer significant. Realizing that the conflicting conclusions arise solely because different models have been used, the editor challenges the author and referee 2 to explain why they have chosen their particular models. Both justify their model by demonstrating an acceptable fit to the data. What should the editor decide? Arguably, the editor will reject the paper on the grounds that the claimed significance depends on an arbitrary choice of model rather than on the actual evidence contained in the data. It is the extent to which inferences of this kind depend on the model, rather than on the data, that we want to look at.

Section 2 looks at two simple textbook examples. Section 2.1 (Example 1) is a 2×4 contingency table giving the frequencies of success/failure tabulated against 4 levels of an ordered categorical covariate. The aim is to estimate the probability of success at a given category. We could just use the saturated model M_0, which makes no modeling assumptions at all about the probabilities in the table, or we could get more accurate estimates by assuming that the probabilities are linked by a logistic model, M say. These give respective confidence intervals CI_{M_0} and CI_M. There are many possible choices for M depending on the structural assumptions made about the ordered categories, and many of these appear to give an acceptable fit to the data as judged by the usual χ^2 test.

Section 2.2 (Example 2) is a multiple regression dataset giving observed values of response y and 13 covariates x. The aim here is to predict the expected value of y for a given set of xs. We could just fit the usual multiple regression model with all 13 covariates, model M_0 say, giving confidence interval CI_{M_0}, or we could try and achieve more accurate estimates by fitting a subset regression M using only some of the x's, giving a shorter confidence interval CI_M. Taking the usual F test as the means of judging the adequacy of different subsets, we find that many subset regressions give an acceptable fit to the dataset. In both examples, we are trying to reduce the number of unknown parameters by fitting sub-models M nested within a more general model M_0.

Although these two examples are very different, they both suggest similar conclusions as far as this paper is concerned. In both examples, we are trying to estimate a single *parameter of interest*, ϕ say, we have a *base model* M_0, and are interested in sub-models M, nested within M_0, which we are assessing by a standard goodness-of-fit test (χ^2 or F). If $\mathcal{M}$ is the set of empirically acceptable models as judged by this test, we look at the overall spread of confidence intervals given by

$$\mathcal{I} = \cup_{M \in \mathcal{M}} \mathrm{CI}_M.$$

In both examples, we find that $\mathcal{I}$ is very much wider than CI_{M_0}, indicating a great deal of variability between CI_Ms for different empirically acceptable models. There are many values of $\phi \in \mathcal{I}$ falling substantially outside CI_{M_0}. This means that if we take the customary naive interpretation of a confidence interval as the set of parameter values that can be considered reasonable in the light of the data, then the uncritical use of goodness of fit as the sole means of assessing models has the perverse consequence of entertaining values of ϕ which would appear reasonable under an empirically acceptable model, but quite unreasonable if we just relied on the basic model M_0 without making any further modelling assumptions. In both examples, we also find many instances of pairs of models in $\mathcal{M}$ which have very similar goodness-of-fit statistics but which give confidence intervals which are disjoint. So, parameter values which are judged reasonable under one model may seem quite unreasonable under another model, even though the two models fit the data equally well. Of course we know that different models will give different inferences, but this degree of inconsistency between apparently sensible models does seem surprising.

Section 3 shows that the qualitative conclusions seen in these two examples hold much more generally, at least asymptotically. In this wider setting, covering many of the simpler problems of practical interest, we continue to have a parameter of interest ϕ, a base model M_0 representing the most general model we are prepared to consider, and a relevant goodness-of-fit statistic. Section 3.3 compares $\mathcal{I}$ with CI_{M_0} and shows that, under standard regularity conditions, $\mathcal{I}$ is wider than CI_{M_0} by a factor of at least $\sqrt{2}$. Section 3.4 examines what this means in terms of significance tests for a given point null hypothesis of the form $H_0 : \phi = \phi_0$, and looks at when we have significance for (a) at least one $M \in \mathcal{M}$, and (b) for all $M \in \mathcal{M}$. Conditions for (a) are extremely weak: Given any value of ϕ_0 there will always be at least one model $M \in \mathcal{M}$ which indicates rejection of the null hypothesis (so the p value is essentially 1) . Condition (b) is clearly much stronger than requiring significance under any single model and, perversely, is considerably stronger even than requiring significance under M_0. (The required p value is much smaller than the p value we would need if we had just used the base model M_0.)

Section 4 looks more carefully at subset selection in multiple regression, again assuming that the aim of the analysis is prediction of the response at ψ, a given vector of the covariates. Now the models M are restricted to linear regressions on subsets of the regressors, a special case of the more general definition of M in Sect. 3. Explicit results show that the bounds obtained in Sect. 3.3 continue to apply for most, but not all, values of ψ. For prediction in regression, we show that the important role of the

parameter of interest is the angle between the vector ψ and the vector of least squares regression coefficients in the full model.

These observations suggest that the fictitious 'editor's dilemma' mentioned at the start of this section is by no means unusual. If the editor believes that there are good scientific reasons for preferring M_1 to M_2, then the conclusion indicated by M_1 would prevail (or the other way round). But if the editor views the matter as purely empirical and accepts that in the light of the data there seems little to choose between them, then there is good reason for rejecting the paper on the grounds that significance depends entirely on an arbitrary choice of model. If several referees, using different empirically acceptable models, all agree that the treatment is significant, then the editor would no doubt accept the paper. But this would amount to condition (b) mentioned above, namely that the null value ϕ_0 would need to be outside the composite interval $\mathcal{I}$. Section 3.3 shows that condition (b) is an absurdly strong requirement; it would be much better to abandon all additional modeling assumptions and just use M_0 to allow the data to speak for themselves. Conversely, as shown in Sect. 3.4, the fact that the author of the paper has found an empirically acceptable model M_1 leading to significance (condition (a)) tells us almost nothing about the truth or otherwise of the null hypothesis. All these raise questions about the customary use of goodness of fit as the sole (or even the main) arbiter for model choice. Model diagnostics are clearly useful if our aim is data description, but they are not enough if our aim is inference focussed on some given parameter of interest. These and other points for discussion are summarized in the concluding Sect. 5.

There has been much comment on what has been described as the 'reproducibility crisis' in the scientific literature. Wadman (2013) reports that in a high proportion of cases, independent researchers have been unable to verify claims made in published research in the biomedical sciences, recalling the title of Ioannidis (2005), 'Why most published research findings are false.' The inadequate reporting of data analysis is frequently cited. Simmons et al. (2011) claim that 'undisclosed flexibility' in published papers in psychology 'allows presenting anything as significant.' Given that a statistical method depends on the model M on which it is based, 'inadequate reporting of data analysis' implies the inadequate recognition of M and the reason that this particular M has been selected.

Of course there is nothing new in pointing out problems caused by ignoring model uncertainty. Many papers show that when a model is selected to be the 'best' model according to some specified selection criterion, the actual coverage of the resulting confidence interval can be noticeably less than the nominal level. Bootstrap methods can be used to estimate marginal properties of post-selection estimates (Efron 2014). The large literature on subset selection in regression (Miller 2002) shows, however, that the choice of model selection criterion can be critical, with different methods resulting in quite different subsets and sometimes sharply different predictions. There is also a large literature on Bayesian model averaging, using a two-stage model covering uncertainty both within and between a given set of candidate models. Hoeting et al. (1999) give an accessible introduction. However, the approach is not universally accepted: Efron (2014) refers to the 'intimidating amount of prior knowledge' required, and Cox (1995) argues that inferences from different models are of interest in their own right. Hjort and Claeskens (2003) discuss a wide-ranging methodology

for frequentist approaches to model averaging. Claeskens and Hjort (2008) provide a good review of research in this area, including some of the information-based criteria which are often recommended for model selection. The problem of choosing subsets of predictors in linear regression is taken as a key example in much of this literature.

Model uncertainty is also discussed in many papers outside the usual statistical literature. Despite the very different terminology, many papers within the machine learning literature relate to these topics: Langford (2005) provides an accessible way into this literature by discussing some machine learning approaches to assessing error rates in binary classification. Of the many related papers in the econometrics literature, Potscher (1991) and Leeb and Potscher (2005) question the commonly held assumption that if a selection criterion is consistent then, for large enough sample sizes, the model selection process can safely be ignored. Non-uniformity of convergence near model boundaries means that asymptotic distributions can exhibit features that are not even approximately true, however large the sample size might be. These papers also raise doubts about the validity of some of the current model selection proposals in the literature.

The technical nature of most of these papers, however, means that the problem of model uncertainty is almost never mentioned in elementary statistics textbooks or courses, and so this literature is largely inaccessible to most users of statistical methods. All too often, model-based inferences, in medical papers for example, are reported as if they are *the* definitive conclusion to be drawn from the data. The substantial variability between model-based inferences using the same data suggests that the role of the model is crucial and needs to be acknowledged much more widely. Demonstrating a reasonable fit to the data is a sensible requirement but is not enough: The model also needs to be seen as sensible in the light of the scientific context.

2 Two examples

We illustrate the variability of inference over different empirically acceptable models by looking at two simple textbook data sets, a 2×4 contingency table from Everitt (1977) and a multiple regression dataset from Royston and Sauerbrei (2008).

2.1 Example 1: modeling a contingency table

The frequencies in Table 1 record the incidence of heart disease (D) against blood pressure (BP), grouped into four ordered categories from low (1) to high (4). Suppose we are interested in assessing the risk at a particular blood pressure category, say level 3. Measuring probability on the logit scale defines the parameter of interest in this case to be

$$\phi = \text{logit}\{\text{P}(D|\text{BP} = 3)\}. \tag{1}$$

The simplest inference for the logit ϕ is just to take the data in the third category as a binomial sample with 20 cases out of $n = 224$ trials. With the continuity correction

Table 1 *Example 1*: heart disease (D) and blood pressure (BP)

	BP = 1	BP = 2	BP = 3	BP = 4
D	20	28	20	24
$\bar{D}$	388	527	204	118
	408	555	224	142

discussed by Cox (1970) of adding 0.5 to both the number of successes and the number of failures, and assuming normality of empirical logits, gives the bias-adjusted estimate and 95% confidence interval

$$\hat{\phi} = -2.30\ ,\ \text{CI} = (-2.75, -1.85). \tag{2}$$

This is the inference from the saturated model, making no parametric modelling assumptions about the probabilities underlying the contingency table.

The data show a steady rise in risk as we move from the lowest to the highest category, suggesting that we would get more accurate estimates if we fitted a model to the complete dataset. A conventional choice might be logistic regression using equally spaced numerical scores for the ordered categories,

$$\text{logit}\{P(D|\text{BP} = i)\} = \alpha + \beta x_i, \tag{3}$$

with the x_is given by $x = (1, 2, 3, 4)$. Estimating (α, β) by maximum likelihood in the usual way gives the fitted probabilities and hence the expected frequencies of the entries in the contingency table (assuming the column totals are fixed). This gives the chi-squared test of goodness of fit as $\chi^2 = 2.99$ on two degrees of freedom, indicating a very acceptable fit. The corresponding estimate and confidence interval for ϕ are

$$\hat{\phi} = -2.24\ ,\ \text{CI} = (-2.48, -2.00). \tag{4}$$

The estimates are quite similar but, as expected, the model-based confidence interval in (4) is much narrower than in (2).

The assumption of a logistic model with equally spaced x_is is clearly arbitrary. Arguably, all we could safely assume a priori is that we have four probabilities which increase as we go from the low to the high category. However, any increasing sequence of probabilities can still be written in the form (3) for some increasing sequence x_i. There is no loss of generality if we linearly transform the x_is so that $x_1 = 0$ and $x_4 = 1$, giving $x = (0, a, b, 1)$ with

$$0 \le a \le b \le 1. \tag{5}$$

Each fixed choice of (a, b) gives a different logistic model with its own expected frequencies, chi-squared statistic $\chi^2(a, b)$ on two degrees of freedom, maximum likelihood estimate $\hat{\phi}(a, b)$, and asymptotic 95% confidence limits $\{\hat{\phi}^{(L)}(a, b), \hat{\phi}^{(U)}(a, b)\}$. The previous results (4) are for $a = 1/3$ and $b = 2/3$.

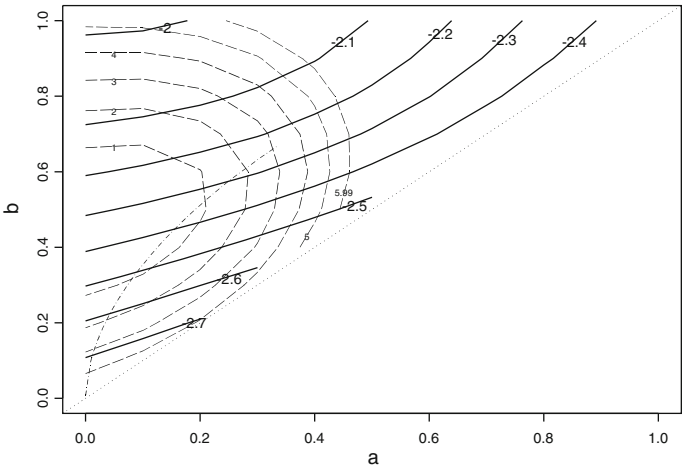

Fig. 1 *Example 1*: Contours of chi-squared (dashed lines) and of estimated values of ϕ (solid lines) for logistic models with scores $(0, a, b, 1)$. The dotted-dashed line corresponds to geometric scores with $x_i \propto \nu^i$ with $\nu > 1$

The range of inferences for different choices of (a, b) is illustrated in Fig. 1. The dashed lines are contours of $\chi^2(a, b)$, plotted in the triangular region (5). The area to the left of the contour labeled 5.99 (the dashed contour furthest to the right) is the region of values of (a, b) which give empirically acceptable models as judged by the chi-squared test at the 5% level. This region includes the equally spaced point $(1/3, 2/3)$, corresponding to an arithmetic progression of the category scores. An increasing geometric progression of scores might seem equally plausible in the context of the data, in which case $x_i \propto \nu^i$ for some $\nu > 1$. The dotted-dashed line on Fig. 1 corresponds to all possible geometric scores with $\nu \geq 1$. The solid lines in Fig. 1 are contours of the estimates $\hat{\phi}(a, b)$. Over the region of empirically acceptable fit, the maximum likelihood estimate of ϕ can take any value between about -2.7 and -2.0, a range considerably wider than the confidence interval (4), and almost as wide as the saturated confidence interval (2). The region also includes the origin in Fig. 1, the limiting geometric model as $\nu \to \infty$, for which

$$\hat{\phi} = -2.80\,, \ \mathrm{CI} = (-3.04, -2.56). \tag{6}$$

When $a = b = 0$ the four logit probabilities in (3) are $(\alpha, \alpha, \alpha, \alpha + \beta)$, suggesting a base line risk which only increases for the largest blood pressure category. Note that most values of ϕ in this confidence interval lie outside the saturated confidence interval (2), and *all* of them lie completely outside the equally spaced confidence interval (4). This indicates a sharp difference between the inferences resulting from assuming equally spaced scores and geometrically spaced scores with a large common ratio.

Figure 2 takes a uniform random sample of points (a, b) from the region (5) with $\chi^2(a, b) \leq 5.99$, and plots the corresponding confidence intervals (vertical line segments) against the chi-squared statistic. Two pairs of horizontal lines are shown. The

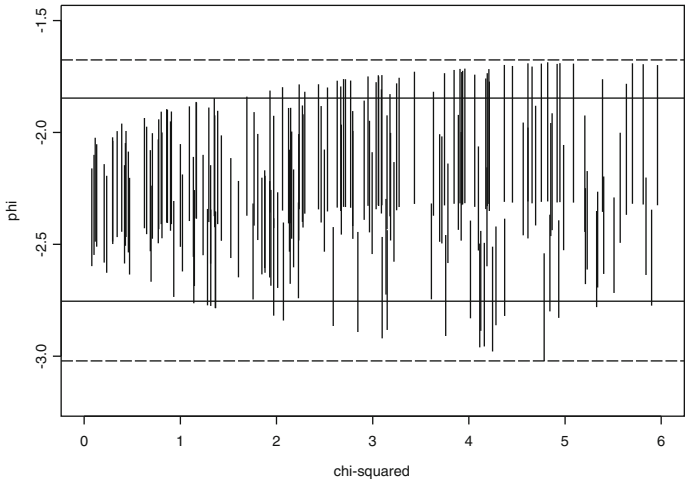

Fig. 2 *Example 1*: Confidence intervals for empirically acceptable models (vertical lines) plotted against values of χ^2_2. Horizontal lines indicate saturated and outer confidence limits

solid lines indicate the saturated confidence limits (2), and the dashed lines indicate the extremes of the model-based confidence limits for $\chi^2 \leq 5.99$, the outer limits

$$\mathcal{I} = \left\{ \min_{\chi^2(a,b)\leq 5.99} \hat{\phi}^{(L)}(a,b) \, , \; \max_{\chi^2(a,b)\leq 5.99} \hat{\phi}^{(U)}(a,b) \right\} = (-3.04, -1.68). \quad (7)$$

This is the range of values of ϕ that are included in the confidence interval for at least one empirically acceptable choice of (a, b) in (5).

Looking at the vertical line segments in Fig. 2 suggests:

- model-based confidence intervals are all shorter than the saturated interval (2);
- there is very considerable variation between confidence intervals for different empirically acceptable models;
- the outer limits $\mathcal{I}$ are considerably wider than the saturated limits (2);
- there are many pairs of empirically acceptable models that give disjoint confidence intervals, including pairs which fit the data equally well as judged by χ^2.

These four comments will be echoed in later sections of the paper, suggesting that they reflect general properties of empirically acceptable models and are not just special features of this particular example.

2.2 Example 2: subsets in multiple regression

One of the examples used in the regression text Royston and Sauerbrei (2008) is a multiple regression dataset in which, for each of 252 male subjects, we have data on percentage body fat (response variable y) and a vector of 13 covariates (x) giving age and weight plus eleven separate body measurements. The data can be downloaded as the file `edu.bodyfat` from the website accompanying Royston and Sauerbrei

(2008). We follow these authors by omitting one aberrant observation (case 39), leaving $n = 251$ observations for analysis. These data were originally discussed in Penrose et al. (1985) and also used as an example in Hoeting et al. (1999).

We assume that these data follow the multiple regression model

$$y = \theta_0 + \theta^{\mathrm{T}} x + \sigma\epsilon,$$

where ϵ is a standard normal residual. Table 2.3 in Royston and Sauerbrei (2008) lists the 13 least squares estimates of θ and their standard errors, showing that only 3 of these are significantly different from zero. This suggests that the model can be simplified by choosing a subset of the x_is and assuming that the remaining θ_is are zero. Many different ways of selecting subsets have been suggested in the literature, so that the choice of any particular subset selection method may seem rather arbitrary. Some of these selection techniques are reviewed in Royston and Sauerbrei (2008) and illustrated on this dataset. The aim of the analysis is explicitly stated in Royston and Sauerbrei (2008, p. 36) as 'to predict the percentage of body fat from the 13 predictors.' So in this example, we define the parameter of interest to be

$$\phi = \mathrm{E}(y|x = \xi) \tag{8}$$

for some given vector ξ of covariates.

With 13 covariates, there are $2^{13} - 1 = 8191$ non-null subsets, each of which can be tested using the F test to compare the fit of the subset regression with the fit of the full regression. For these data, it turns out that 1473 of these subsets pass the F test at the 5% level. Assume that subset S includes k_S covariates and gives confidence interval CI_S for ϕ. If F_S is the F statistic for this subset then F_S is on $(13-k_S, 251-13-1 = 237)$ degrees of freedom. We are interested in the variability of CI_S over those 1473 subsets for which F_S is less than its corresponding null percentage point.

Suppose, for instance, that we want to estimate ϕ in (8) with ξ equal to the first observed value of x in the dataset. Figure 3 takes a random sample of 100 subsets from the 1473 empirically acceptable subsets, and plots the corresponding confidence intervals CI_S (vertical line segments) against a measure of how well subset S fits the data. Since each F_S has different degrees of freedom, we bring them onto a common scale by transforming F_S monotonically into an equivalent value of χ^2 on two degrees of freedom, namely

$$\chi^2_2 = F_1^{-1}\{F_2(F_S)\},$$

where F_1 and F_2 are the respective cumulative distribution functions of χ^2 on two degrees of freedom, and F on $(13 - k_S, 237)$ degrees of freedom. This is simply a technical device to ensure that the horizontal axis of Fig. 3 is directly comparable to that of Fig. 2 in the first example. The horizontal solid lines in Fig. 3 give the corresponding confidence limits from the full regression (therefore matching the vertical line segment shown at $\chi^2 = 0$) and the horizontal dashed lines are the outer extremes of CI_S over all 1473 empirically acceptable subset regressions. If we draw an analogy between the full regression in Example 2 and the saturated model in Example 1, and between the

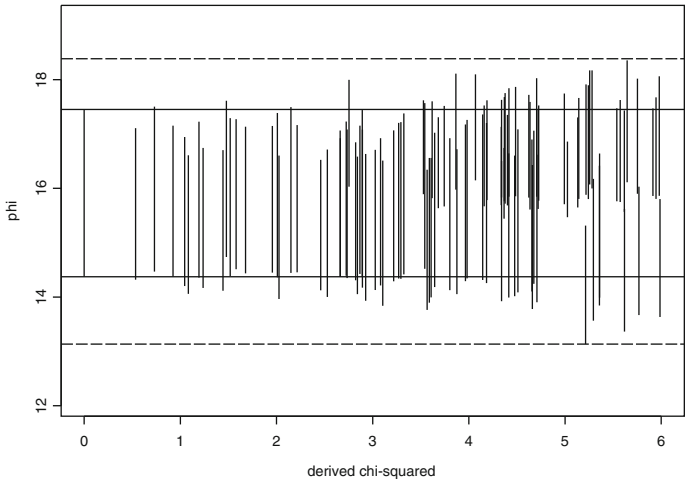

Fig. 3 *Example 2*: Confidence intervals for empirically acceptable subset regression models (vertical lines) plotted against equivalent values of χ^2_2. Horizontal lines indicate confidence limits for the full regression (solid lines), and outer confidence limits over all empirically acceptable subsets (dashed lines)

choice of subset S in Example 2 and the choice of (a, b) in Example 1, we see that there is a striking similarity between Figs. 2 and 3. The comments listed at the end of Sect. 2.1 apply equally well here.

For these data, we find that all empirically acceptable subsets have $k_S \geq 4$ and include the covariates x_6 (abdominal circumference) and x_{13} (wrist circumference), the two most significant covariates in the full regression. We get two empirically acceptable four-covariate subsets by adding x_5 (chest circumference) and then either x_1 (age) or x_3 (height). It turns out that this is an example of two empirically acceptable models which give disjoint confidence intervals, subset (x_1, x_5, x_6, x_{13}) giving $\text{CI}_S = (13.67, 15.84)$, and subset (x_3, x_5, x_6, x_{13}) giving $\text{CI}_S = (15.98, 17.45)$. Given that one wants to choose a subset with the smallest acceptable number of covariates, it is difficult to see that there would be any substantive reason for preferring x_1 or x_3 as the fourth covariate when x_5, x_6 and x_{13} are already included, and so the choice between these two models seems pretty arbitrary. However, the fact that the confidence intervals are disjoint shows that the predictions resulting from these two models differ sharply. The same point also arises with model averaging, where model uncertainty translates into uncertainty about the prior model probabilities. If we assign prior probabilities ρ and $1 - \rho$ to these two subsets, we can see how the 95% highest posterior interval for ϕ depends on ρ. Following the approximation in Draper (1995), section 5.4, these are the same as the disjoint confidence intervals when $\rho = 0$ and 1, and for $\rho = \frac{1}{2}$ we get $(13.87, 17.52)$, quite close to the set union of the two. Using the same data, Hoeting et al. (1999) makes the conventional assumption of a uniform prior distribution, implying that $\rho = \frac{1}{2}$, but offers no substantive reason why this particular value of ρ is appropriate.

3 Asymptotic theory

3.1 Basic setup

As indicated in Sect. 1, and illustrated by both examples in Sect. 2, our basic setup involves four key ingredients: (a) a *base model* M_0, (b) a *model* M, (c) a *goodness-of-fit* statistic $G(x, M)$, and (d) a *parameter of interest* ϕ. In example 2.1 these are, respectively, the saturated model, a model with given values of (a, b), the χ^2 test, and the logit disease risk for the third blood pressure category. In Example 2.2, they are the full multiple regression model, a subset regression, the F test, and the predicted response at a given set of covariates. In this section, we generalize these examples into a wider parametric setting by making the following assumptions.

(a) *Base model* M_0.

We assume that M_0 is a regular parametric model under which observation x has probability density function $f(x, \theta)$ for a vector parameter θ with k components. For simplicity of notation we shall assume that x is continuous. The base model M_0 gives the score vector and expected information matrix

$$s = s(x, \theta) = \frac{\partial}{\partial \theta} \log f(x, \theta) \ , \ I = I(\theta) = -\mathrm{E}\left\{ \frac{\partial^2}{\partial \theta \partial \theta^{\mathrm{T}}} \log f(x, \theta) \right\}. \tag{9}$$

(b) *Model* M.

We assume that M is nested within M_0 in the sense that observation x still has distribution $f(x, \theta)$, but now θ is restricted to a function $\theta = F_M(\theta_M)$, where θ_M has $k_M < k$ parameters. Assuming that $F_M(\theta_M)$ is a twice differentiable function of θ_M, the score vector and expected information matrix for model M are, respectively,

$$\begin{aligned} s_M &= s_M(x, \theta_M) = D_M^{\mathrm{T}} s\{x, F_M(\theta_M)\} \\ I_M &= I_M(\theta_M) = D_M^{\mathrm{T}} I\{F_M(\theta_M)\} D_M, \end{aligned}$$

where D_M is the $k \times k_M$ matrix

$$D_M = D_M(\theta_M) = \frac{\partial \theta}{\partial \theta_M^{\mathrm{T}}} = \frac{\partial F_M(\theta_M)}{\partial \theta_M^{\mathrm{T}}}. \tag{10}$$

(c) *Goodness-of-fit statistic* G.

For given sample data $x = (x_1, x_2, \ldots, x_n)$ and significance level α, we judge the fit of model M relative to the fit of the more general model M_0 by comparing a test statistic $G(x, M)$ to its asymptotic null percentage point $g_\alpha = g_{\alpha, M}$. This is a size α test of the null hypothesis that $\theta = F_M(\theta_M)$ for some θ_M. We assume that G takes the asymptotic form

$$G(x, M) = H\left[n^{\frac{1}{2}} \left\{ \hat{\theta} - F_M\left(\hat{\theta}_M\right) \right\}, F_M\left(\hat{\theta}_M\right) \right] + O_p\left(n^{-\frac{1}{2}}\right), \tag{11}$$

where $\hat{\theta}$ is the MLE of θ under M_0, $\hat{\theta}_M$ is the MLE of θ_M under M, and $H(c,d)$ is a smooth function of (c,d) satisfying $H(0,d) = 0$ and $H(c,d) = H(-c,d)$. For G to be an omnibus test (sensitive to departures in all directions) we assume that if $H(c,d) \le g_\alpha$ then all of the components of vector c are necessarily finite. The factor $n^{\frac{1}{2}}$ in the first argument of (11) is necessary so that the null distribution of G remains finite as $n \to \infty$. In regular models of the kind considered here, the asymptotic form of most goodness-of-fit tests of practical interest can be expressed in the form (11). In both of the examples in Sect. 2, $H(c,d)$ is a positive definite quadratic form in c.

(d) *Parameter of interest* ϕ.

We assume that the parameter of interest is a differentiable scalar function $\phi(\theta)$ of the parameter θ of the base model M_0. If model M is true then $\phi = \phi\{F_M(\theta_M)\}$.

These strong regularity assumptions will hold for many of the simpler problems of practical interest, but clearly not all. In particular, we exclude all cases with non-identifiability problems, such as missing data or hidden confounders.

3.2 Maximum likelihood estimates and asymptotic confidence intervals

The maximum likelihood estimates of ϕ under models M_0 and M are $\phi(\hat{\theta})$ and $\phi(F_M(\hat{\theta}_M))$, respectively. Using standard maximum likelihood asymptotics we can approximate the variances of these estimates, and the covariance between them, in terms of the information matrix I defined in (9), from which we get

$$\mathrm{Var}\left(\hat{\phi} - \hat{\phi}_M\right) = n^{-1}\left(\sigma^2 - \sigma_M^2\right) + o\left(n^{-1}\right), \tag{12}$$

where

$$\sigma^2 = n\mathrm{Var}\left(\hat{\phi}\right) = \frac{\partial\phi}{\partial\theta^{\mathrm{T}}} I^{-1} \frac{\partial\phi}{\partial\theta} + O\left(n^{-\frac{1}{2}}\right) \tag{13}$$

$$\sigma_M^2 = n\mathrm{Var}\left(\hat{\phi}_M\right) = \frac{\partial\phi}{\partial\theta^{\mathrm{T}}} D_M \left(D_M^{\mathrm{T}} I D_M\right)^{-1} D_M^{\mathrm{T}} \frac{\partial\phi}{\partial\theta} + O\left(n^{-\frac{1}{2}}\right) \tag{14}$$

Equation (12) is just the usual analysis of variance identity for linear models when applied to local linear approximations to ϕ and M. An immediate consequence is that $\sigma_M^2 \le \sigma^2$, confirming that model M always gives a shorter asymptotic confidence interval than the base model M_0. We can see this in Figs. 2 and 3 by noting that the lengths of the vertical line segments are all less than the distance between the two horizontal solid lines.

The variances defined in (13) and (14) give the asymptotic confidence intervals of ϕ under models M_0 and M to be

$$\mathrm{CI} = \mathrm{CI}_{M_0} = \left(\hat{\phi} - n^{-\frac{1}{2}} z_\alpha \sigma \ ,\ \hat{\phi} + n^{-\frac{1}{2}} z_\alpha \sigma\right)$$

$$\mathrm{CI}_M = \left(\hat{\phi}_M - n^{-\frac{1}{2}} z_\alpha \sigma_M \ ,\ \hat{\phi}_M + n^{-\frac{1}{2}} z_\alpha \sigma_M\right),$$

where z_α is the $(1-\alpha/2)$ quantile of the standard normal distribution. Also, from (12), the statistic

$$z = z(x, M) = \frac{n^{\frac{1}{2}}\left(\hat{\phi}_M - \hat{\phi}\right)}{\left(\sigma^2 - \sigma_M^2\right)^{\frac{1}{2}}} \tag{15}$$

is asymptotically standard normal under model M. This allows us to rewrite CI_M as

$$\mathrm{CI}_M = \left\{\hat{\phi} + n^{-\frac{1}{2}}\left[\left(\sigma^2 - \sigma_M^2\right)^{\frac{1}{2}} z - z_\alpha \sigma_M\right],\ \hat{\phi} + n^{-\frac{1}{2}}\left[\left(\sigma^2 - \sigma_M^2\right)^{\frac{1}{2}} z + z_\alpha \sigma_M\right]\right\}. \tag{16}$$

This shows that the asymptotic model-based confidence interval for ϕ depends on the model M through just two scalar quantities, the model variance σ_M^2 and $z(x, M)$ reflecting how well model M fits the data. It is easy to show that z in (15) is just the scalar projection of the usual score test for M in the direction given by $\partial\phi/\partial\theta$, and so can be thought of as the log likelihood ratio test most relevant to detecting differences in ϕ.

3.3 Bounds for confidence limits for empirically acceptable models

The expressions in square brackets in (16) are proportional to the differences between the confidence limits under M and the base estimate $\hat{\phi}$. They involve a classic bias/variance compromise: as σ_M^2 decreases the size of the variance term $\pm z_\alpha \sigma_M$ decreases, but the size of the bias term $(\sigma^2 - \sigma_M^2)^{\frac{1}{2}} z$ increases. We naturally prefer a model M which gives a small variance σ_M^2, but we also need to control the size of the bias. In practice, we do this by confirming (or assuming) that our model gives an adequate fit to the data as measured by a goodness-of-fit test. We are thus interested in the size of $z = z(x, M)$ for models M with $G(x, M) \leq g_\alpha$.

We show in Supplementary Appendix A that for any fixed value of $\sigma^* \leq \sigma$, and for sufficiently large n,

$$\sup_M \left\{|z(x, M)| :\ \sigma_M = \sigma^*,\ G(x, M) \leq g_\alpha\right\} > z_\alpha \tag{17}$$

for almost all data vectors x. Thus, if we denote the upper and lower confidence limits of CI_M in (16) by $\mathrm{CI}_M = (\mathrm{CI}_M^{(L)}, \mathrm{CI}_M^{(U)})$, (17) gives

$$\begin{aligned}
\mathcal{I}^{(U)} = \sup_{M: G(x,M) \leq g_\alpha} \mathrm{CI}_M^{(U)} &= \sup_{M: G(x,M) \leq g_\alpha} \left\{\hat{\phi} + n^{-\frac{1}{2}}\left[\left(\sigma^2 - \sigma_M^2\right)^{\frac{1}{2}} z + z_\alpha \sigma_M\right]\right\} \\
&> \sup_{\sigma^{*2} \leq \sigma^2} \left\{\hat{\phi} + n^{-\frac{1}{2}} z_\alpha \left[\left(\sigma^2 - \sigma^{*2}\right)^{\frac{1}{2}} + \sigma^*\right]\right\} \\
&= \hat{\phi} + 2^{\frac{1}{2}} n^{-\frac{1}{2}} z_\alpha \sigma\ .
\end{aligned} \tag{18}$$

The last step follows from the elementary inequality

$$\left(\sigma^2 - \sigma^{*2}\right)^{\frac{1}{2}} + \sigma^* \leq 2^{\frac{1}{2}}\sigma,$$

with equality attained when $\sigma^{*2} = \frac{1}{2}\sigma^2$. Reversing the sign in (18) gives $\mathcal{I}^{(L)}$, the complementary bound for the lower confidence limit. So for large n and almost all x,

$$\mathcal{I} = \left(\mathcal{I}^{(L)}, \mathcal{I}^{(U)}\right) \supset \mathrm{CI}^{(2)} = \left(\hat{\phi} - 2^{\frac{1}{2}} n^{-\frac{1}{2}} z_\alpha \sigma \ , \ \hat{\phi} + 2^{\frac{1}{2}} n^{-\frac{1}{2}} z_\alpha \sigma\right). \qquad (19)$$

The interval $\mathrm{CI}^{(2)}$ on the right-hand side of (19) is the base model confidence interval CI widened by a factor of $\sqrt{2}$ (or doubling the variance). The strict inequality in (17) arises from the assumption that G is an omnibus test, which precludes the possibility that $G = |z|$. But by thinking of $|z|$ as the limit of a sequence of omnibus tests giving increasing weight to the particular direction given by $\phi = \phi(\theta)$, these inequalities can be rewritten as the asymptotic minimax property

$$\left[\sup_G \left\{\inf_{M:G(x,M)\leq g_\alpha} \mathrm{CI}_M^{(L)}\right\} \ , \ \inf_G \left\{\sup_{M:G(x,M)\leq g_\alpha} \mathrm{CI}_M^{(U)}\right\}\right] = \mathrm{CI}^{(2)}.$$

These bounds can be expressed more simply as follows, taking the conventional values $\alpha = 0.05$ and $z_\alpha = 1.96$. Consider any value of ϕ between the limits $\hat{\phi} \pm 2.77 n^{-\frac{1}{2}}\sigma$. Then for any 5% goodness-of-fit test G of the form (11), $\phi \in \mathrm{CI}_M$ for at least one model M which gives an acceptable fit to the data as judged by G.

We have used a very wide definition of models M by allowing $\theta = F_M(\theta_M)$ to be *any* smooth function of a model-based parameter vector θ_M of lower dimension than θ. But in practice, we will want to restrict the choice of models to those that might be considered sensible in the context of the data. The interval $\mathcal{I}$ restricted to the models M being considered may therefore be shorter and no longer satisfy (19). Our two examples, however, suggest that confidence intervals for empirically acceptable models of practical interest can still reach considerably beyond the limits of the base interval CI. In Example 1, modeling assumptions only involve disease probabilities within each column of the contingency table and leave the column totals as free parameters, and even then, only models with monotonic risks are considered. In this case, the extremes of the confidence limits for ϕ among models that pass the χ^2 test are $(-3.02, -1.68)$, which is much wider than $\mathrm{CI} = (-2.75, -1.85)$, and quite similar to the doubled-variance interval $\mathrm{CI}^{(2)} = (-2.94, -1.66)$. In Example 2, the only models considered were subset regressions, but again we find a similar pattern for predicting the expected response at the particular ξ we considered. The extremes of the confidence limits resulting from subset regressions that pass the F test are (13.14, 18.39), much wider than the full regression confidence interval $\mathrm{CI} = (14.37, 17.46)$, and also wider than the doubled-variance interval $\mathrm{CI}^{(2)} = (13.74, 18.09)$.

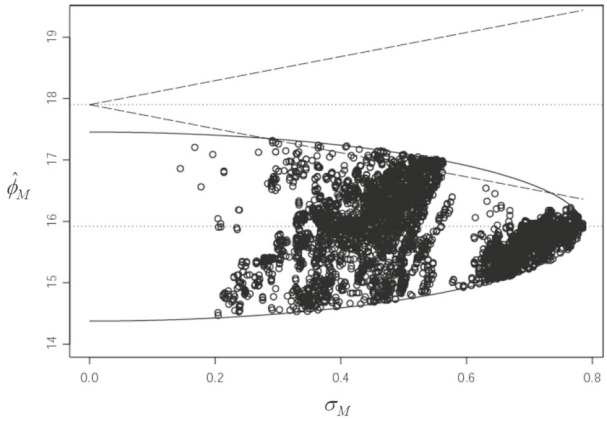

Fig. 4 *Example 2*: Values of $\hat{\phi}_M$ plotted against σ_M for all subset regressions with $|z| \leq 1.96$. The lines relate to the discussion in Sect. 3.4

3.4 Significance tests

Suppose we wish to test the hypothesis $H_0 : \phi = \phi_0$ for some null value ϕ_0 of interest. Then, for any given model M, the test which rejects H_0 if and only if $\phi_0 \notin \mathrm{CI}_M$ is an asymptotic significance test of level α. The above discussion of asymptotic confidence intervals can therefore be immediately recast in terms of asymptotic significance tests.

The acceptance region of the test corresponding to CI_M is

$$|\hat{\phi}_M - \phi_0| \leq n^{-\frac{1}{2}} z_\alpha \sigma_M, \tag{20}$$

and, if the test statistic z in (15) is used as a measure of model fit, then model M is empirically acceptable if

$$|\hat{\phi}_M - \hat{\phi}| \leq n^{-\frac{1}{2}} z_\alpha (\sigma^2 - \sigma_M^2)^{\frac{1}{2}}. \tag{21}$$

These two regions are illustrated in Fig. 4, which plots values of $\hat{\phi}_M$ against σ_M for $0 \leq \sigma_M \leq \sigma$. For this illustration, we have assumed $\alpha = 0.05$ and taken the values $n = 251$, $\hat{\phi} = 15.92$ and $\sigma = 0.79$ from Example 2 of Sect. 2.2. The semi-elliptical region bounded by the solid line consists of the values of $(\sigma_M, \hat{\phi}_M)$ within (21), and the triangular region bounded by the dashed lines consist of the points in (20). The apex of the triangle locates the null value ϕ_0, illustrated here for the arbitrarily chosen value $\phi_0 = 17.92$. The scatter of points shown in Fig. 4 corresponds to all the values of $(\sigma_M, \hat{\phi}_M)$ for subset regressions in Example 2 with $|z(x, M)| \leq 1.96$.

The intersection of the two regions in Fig. 4 delineates those models that both fit the data in the sense that $|z| \leq z_\alpha$ and also lead to acceptance of H_0. As ϕ_0 moves away from the base model estimate $\hat{\phi}$, or as the dashed lines in Fig. 4 move upwards, this intersection becomes smaller and vanishes when the lower of these lines just touches the upper ellipse. This happens at the point $(\sigma_M = 2^{-\frac{1}{2}}\sigma, \hat{\phi}_M = \hat{\phi} + n^{-\frac{1}{2}} z_\alpha \sigma_M)$,

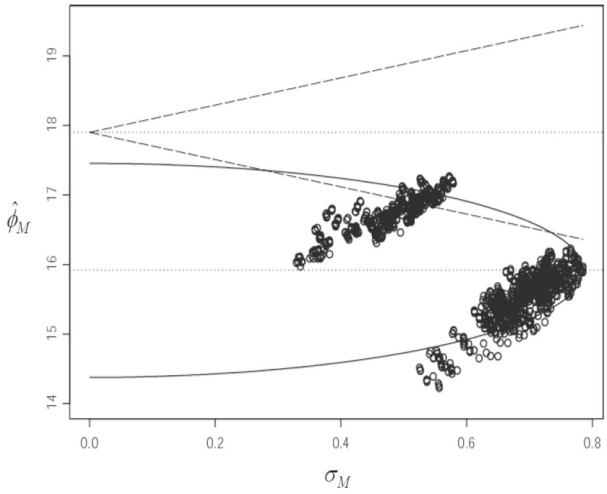

Fig. 5 *Example 2*: Values of $\hat{\phi}_M$ plotted against σ_M for all subset regressions that give an acceptable fit as judged by the F test. The lines relate to the discussion in Sect. 3.4

at which ϕ_0 just falls on the boundary of the doubled-variance confidence interval $\mathrm{CI}^{(2)}$. It follows that, if and only if $\phi_0 \notin \mathrm{CI}^{(2)}$, H_0 will be rejected by all models with $|z(x, M)| \leq z_\alpha$. This means that requiring rejection for all such models is a very strong requirement since, even if the null hypothesis is true and we only assume the base model M_0, the actual asymptotic significance level would be $2\Phi(-2^{\frac{1}{2}} z_\alpha) = 0.0055$ if $\alpha = 0.05$. Evidently, it would be better to abandon the search for empirically acceptable models M and simply use the base model to retain the nominal significance level. Note also that, whatever the value of ϕ_0, the semi-elliptical region in Fig. 4 can never lie wholly within the triangle, which means that there is no null hypothesis that would be judged acceptable by all models with $|z| \leq z_\alpha$. Equivalently, whatever the value of ϕ_0, there will always be at least one model M with $|z(x, M)| \leq z_\alpha$ which leads to rejection of H_0.

As discussed in Sect. 3.3, when fit is judged by an omnibus test, there may be empirically acceptable models M with $|z(x, M)| > z_\alpha$. The points $(\sigma_M, \hat{\phi}_M)$ for such models will then lie outside the ellipse shown in Fig. 5. Figure 5 is the same as Fig. 4, but now shows the points $(\sigma_M, \hat{\phi}_M)$ for all the subset regressions in Example 2 that give an acceptable overall fit as judged by the omnibus F test. The clusters of points in these two graphs are quite different. There are many subsets included in Fig. 4 but excluded from Fig. 5. These are the subsets which give estimates of ϕ close to $\hat{\phi}$ but give a poor fit to the regression as a whole. Conversely, there are subsets in Fig. 5 but not in Fig. 4. These correspond to the points falling outside the ellipse, subsets giving an acceptable overall fit but only at the expense of a poorer fit in the particular direction ϕ.

Models deemed empirically acceptable by an omnibus test, but giving points outside the ellipse, may lead to acceptance of H_0 even though $\phi_0 \notin \mathrm{CI}^{(2)}$. Using the numerical results for Example 2 quoted in Sect. 2.2, the condition for H_0 to be rejected by all

models that pass the F test is that $\phi_0 \notin \mathcal{I} = (13.14, 18.39)$. This is even more stringent than requiring that $\phi_0 \notin \text{CI}^{(2)} = (13.74, 18.09)$. Requiring rejection of H_0 by all empirically acceptable models is a very strong requirement, much stronger than merely to require rejection under the base model, which for this example means that $\phi_0 \notin \text{CI} = (14.47, 17.46)$.

The earlier comment that there is no null hypothesis that would be accepted by all models with $|z| \leq z_\alpha$, continues to hold when model fit is judged by an omnibus test. In both Figs. 2 and 3, for example, we see that there is no value of ϕ that is covered by all the vertical line segments shown on the graph. Equivalently, for any fixed value of ϕ, there will always be at least one vertical line segment that does not cover ϕ. This holds for any goodness-of-fit test of the form (11) and for any value of ϕ_0 : There will always be at least one empirically acceptable model that indicates rejection of H_0. So, in the 'editor's dilemma' of Sect. 1, if the *only* reason given by the author for using model M_1 is that it gives an acceptable fit to the data, then the editor can reject the author's analysis as worthless.

4 Example 2 revisited: prediction from subset regression

We have used Example 2 (Sect. 2.2) several times to illustrate our discussion. However, as remarked earlier, confining attention to subset regressions represents only a special case of the general definition of M, and our completely arbitrary use of the first observation in the dataset to define ϕ means that we have ignored the important role of the choice of the covariate vector at which we wish to predict. In this section, we take a more careful look at the problem of prediction using subsets in multiple regression.

4.1 Notation and setup

Using the familiar notation for regression, let M_0 be the standard linear regression model for the regression of response variable y on covariate x, a vector of k explanatory variables. Little is lost by ignoring the intercept, so assume from now on that model M_0 is

$$M_0 \ : \ y|x \sim N\left(\theta^{\mathrm{T}} x, \sigma^2\right). \tag{22}$$

We have n independent observations (y_i, x_i), $i = 1, 2, \ldots, n$. Our goal is to predict y for a given vector ξ of covariates, for which we need to estimate

$$\phi = \mathrm{E}(y|x = \xi) = \theta^{\mathrm{T}} \xi. \tag{23}$$

There is no loss of generality if we assume that ξ is scaled so that $\xi^{\mathrm{T}} \xi = 1$. For the data in Sect. 2.2, $k = 13$, $n = 251$ and ξ was taken as the (centered and scaled) vector x observed for the first subject in the dataset.

With a relatively large number of covariates, it is standard practice to try and simplify the model by reducing its dimension. This might be done by selecting a subset of the

covariates in the form that they have been measured. For the data in Example 2, Figure 2.1 in Royston and Sauerbrei (2008, p. 38) shows that backward elimination results in a subset of only 4 covariates which gives fitted values of y quite close to those of the full regression with all 13 covariates. In practice, however, we may wish to first transform the covariates into more meaningful predictors. In the example, the body measurements x_4 to x_{13} are clearly related to the basic variables x_1 to x_3 of age, weight and height, and so for predicting percentage body fat it might be more meaningful to first transform to age-corrected body measurements, or to measures corrected for their dependence on some or all of x_1, x_2 and x_3. Essentially, this would mean replacing a measure by its residual from an appropriate regression among the covariates. This suggest that subset regressions could usefully be extended to regressions of y on vectors of linearly transformed covariates.

We are thus interested in models M given by

$$M: \ y|x \sim N\left(\theta_M^{\mathrm{T}} x_M, \sigma^2\right), \tag{24}$$

where $x_M = A_M x$ and A_M is a $k_M \times k$ matrix with $k_M < k$. A subset regression is the special case when A_M is a matrix of zeros and ones, with one 1 in each of its k_M rows. Although the coefficients and dimension of A_M may have been estimated from the same data, we continue to be interested in the standard practice of assuming that each submodel is fixed, thus treating A_M as if it had been fixed in advance. Note that if model M is true, then both (22) and (24) must have the same residual variance σ^2. The omnibus F test is just the empirical check of this assumption.

In the general notation of Sect. 3, (24) asserts that $\theta = F_M(\theta_M) = A_M^{\mathrm{T}}\theta_M$ and (23) asserts that $\phi = \phi(\theta) = \xi^{\mathrm{T}}\theta$, so in this case both $F_M(\theta_M)$ and $\phi(\theta)$ are linear functions. We are interested in the models generated by all possible $k_M \times k$ matrices A_M of full row rank with $1 \le k_M \le k-1$. Let X be the full $n \times k$ design matrix of the observed covariate vectors. Allowing for all possible matrices A_M means that there is no loss of generality if we assume that x has been linearly transformed to orthonormal form. Similarly, model M is invariant under linear transformations of x_M. This means that we can assume from now on that

$$X^{\mathrm{T}}X = nI_k \ , \ A_M A_M^{\mathrm{T}} = I_{k_M}.$$

The least squares estimate of θ and its variance under the base model M_0 are then simply

$$\hat{\theta} = \frac{1}{n}X^{\mathrm{T}}y \ , \ \mathrm{Var}\left(\hat{\theta}\right) = \frac{\sigma^2}{n}I_k,$$

where $y = (y_1, y_2, \ldots, y_n)$ is the vector of observed values of the response variable. Similarly, for model M we have

$$\hat{\theta}_M = \left(A_M X^{\mathrm{T}} X A_M^{\mathrm{T}}\right)^{-1} A_M X^{\mathrm{T}} y = A_M \hat{\theta} \ , \ \mathrm{Var}\left(\hat{\theta}_M\right) = \frac{\sigma^2}{n}P, \tag{25}$$

where

$$P = A_M^{\mathrm{T}} A_M.$$

Note that for the $k \times k$ matrix P we have $PP = P$ and so P is idempotent for all A_M.

In the notation of Sect. 3.2, it follows immediately that

$$\hat{\phi} = \hat{\theta}^{\mathrm{T}} \xi \ , \ \mathrm{Var}\left(\hat{\phi}\right) = \frac{\sigma^2}{n}. \tag{26}$$

Comparing (26) with (13) confirms that there is no conflict of notation in using σ^2 both for the residual variance and for its earlier definition as $n\mathrm{Var}(\hat{\phi})$. Also

$$\hat{\phi}_M = \hat{\theta}^{\mathrm{T}} P \xi \ , \ \sigma_M^2 = n\mathrm{Var}\left(\hat{\phi}_M\right) = \sigma^2 \xi^{\mathrm{T}} P \xi. \tag{27}$$

The earlier inequality $\sigma_M^2 \leq \sigma^2$ is an immediate consequence of the fact that P is idempotent and $\xi^{\mathrm{T}} \xi = 1$.

4.2 Confidence intervals for ϕ

From (26) and (27), the asymptotic confidence intervals for ϕ under models M_0 and M are

$$\begin{aligned} \mathrm{CI} &= \left\{\hat{\theta}^{\mathrm{T}} \xi - \delta \ , \ \hat{\theta}^{\mathrm{T}} \xi + \delta\right\} \\ \mathrm{CI}_M &= \left\{\hat{\theta}^{\mathrm{T}} P \xi - \delta \left(\xi^{\mathrm{T}} P \xi\right)^{\frac{1}{2}} \ , \ \hat{\theta}^{\mathrm{T}} P \xi + \delta \left(\xi^{\mathrm{T}} P \xi\right)^{\frac{1}{2}}\right\}, \end{aligned} \tag{28}$$

where

$$\delta = \frac{\sigma}{\sqrt{n}} z_\alpha.$$

As in the general case, we also have

$$n\mathrm{Var}\left(\hat{\phi}_M - \hat{\phi}\right) = \sigma^2 - \sigma_M^2,$$

and so $\hat{\phi}_M$ and $\hat{\phi}$ are not significantly different at level α if

$$\left|\hat{\theta}^{\mathrm{T}} (I_k - P) \xi\right| \leq \delta \left\{\xi^{\mathrm{T}} (I_k - P) \xi\right\}^{\frac{1}{2}}. \tag{29}$$

In the previous notation, this is the condition $|z| \leq z_\alpha$. We showed in Sect. 3.3 that, with the more general definition of M, the maximum upper limit of CI_M over empirically acceptable models was least when models are accepted if $|z| \leq z_\alpha$. So to investigate

what happens when M is restricted to the linear form assumed here, we are interested in the extremes of CI_M when $A_M \in \mathcal{A}$, where

$$\mathcal{A} = \bigcup_{k_M=1}^{k-1} \left\{ A_M : A_M A_M^{\mathrm{T}} = I_{k_M} \ , \ \left| \hat{\theta}^{\mathrm{T}} \left(I_k - A_M^{\mathrm{T}} A_M \right) \xi \right| \le \delta \left\{ \xi^{\mathrm{T}} \left(I_k - A_M^{\mathrm{T}} A_M \right) \xi \right\}^{\frac{1}{2}} \right\}.$$

It turns out that the crucial property of ξ is the angle between ξ and $\hat{\theta}$, measured by the correlation

$$r = \frac{\hat{\theta}^{\mathrm{T}} \xi}{\left(\hat{\theta}^{\mathrm{T}} \hat{\theta} \right)^{\frac{1}{2}}}.$$

We show in Supplementary Appendix B that, for sufficiently large n,

$$\sup_{A_M \in \mathcal{A}} \mathrm{CI}_M^{(U)} = \begin{cases} \hat{\theta}^{\mathrm{T}} \xi + 2^{\frac{1}{2}} \delta & \text{if } |r| \le 2^{-\frac{1}{2}} \\ \hat{\theta}^{\mathrm{T}} \xi + \delta \{ (1 - r^2)^{\frac{1}{2}} + |r| \} & \text{if } |r| > 2^{-\frac{1}{2}} \end{cases} . \tag{30}$$

This can be compared with $\hat{\theta}^{\mathrm{T}} \xi + 2^{\frac{1}{2}} \delta$, the upper bound for the general case in (18). The corresponding results for the lower confidence limits are exactly the same with the sign of δ reversed. For smaller values of r, meaning that ξ is not too co-linear with $\hat{\theta}$, these bounds are exactly the same. For values of r closer to ± 1, $\hat{\phi}_M$ and σ_M^2 become more closely linked and the outer limits are less extreme. When $r = 1, \xi = \hat{\theta} / (\hat{\theta}^{\mathrm{T}} \hat{\theta})^{\frac{1}{2}}$, and so the confidence limits CI_M can be written

$$\left(\hat{\theta}^{\mathrm{T}} \hat{\theta} \right)^{\frac{1}{2}} B \pm \delta B^{\frac{1}{2}},$$

where $B = (\hat{\theta}^{\mathrm{T}} P \hat{\theta}) / (\hat{\theta}^{\mathrm{T}} \hat{\theta})$. Since P is idempotent, $B \le 1$, confirming that in this case $\mathrm{CI}_M \subseteq \mathrm{CI}$ as indicated in (30).

The extremes of confidence limits are illustrated in Fig. 6, which shows values of

$$\frac{\max_{\phi \in \mathrm{CI}_{(G)}} \left| \phi - \hat{\phi} \right|}{\max_{\phi \in \mathrm{CI}} \left| \phi - \hat{\phi} \right|} \tag{31}$$

plotted against r for $0 \le r \le 1$. The value of (31) depends on ξ, the vector of covariates for which the prediction confidence intervals are being calculated, and the definition of the class $\mathrm{CI}_{(G)}$ of confidence intervals. The plotted points correspond to the data in Example 2, taking ξ as a weighted combination of the specific ξ used in Sect. 2.2, and the unique vector $\xi \propto X^{\mathrm{T}} X \hat{\beta}$ for which $r = 1$. We have chosen the weights to generate a sequence of vectors ξ with values of r ranging from 0 to 1. For each ξ, we define $\mathrm{CI}_{(G)}$ to be the set union of all confidence intervals CI_M for subset regressions selected by the goodness-of-fit statistic G. The points marked O are when G is the usual F test, the points marked X are when G is taken as $|z|$, i.e., when subsets are accepted if they satisfy (29). The horizontal dotted lines indicate the values $\sqrt{2}$, when

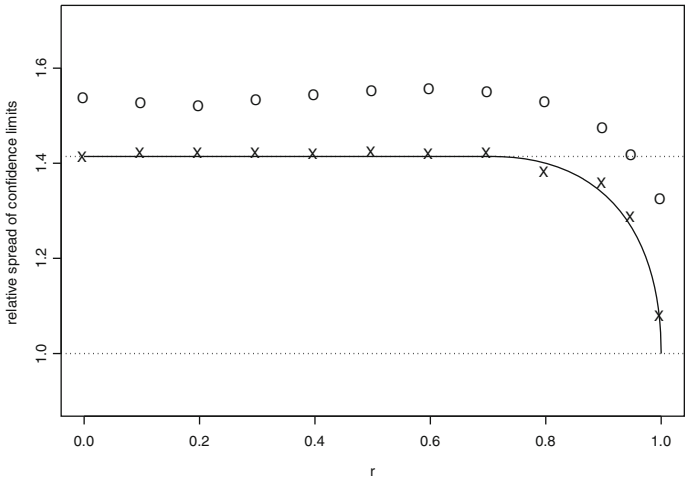

Fig. 6 *Example 2* : Illustrating the relative extremes of subset confidence intervals as defined in (31). Points marked X are for subsets with $|z| \leq 1.96$, points marked O are for subsets that give an acceptable fit as judged by the F test. The curve shows the corresponding asymptotic limits in Eq. (30)

$\mathrm{CI}_{(G)} = \mathrm{CI}^{(2)}$, and 1, when $\mathrm{CI}_{(G)} = \mathrm{CI}$. The curve in Fig. 6 is the multiple of δ in (30), the asymptotic value of (31) when $\mathrm{CI}_{(G)}$ is defined as the union of the more general confidence intervals CI_M in (28). The points marked X are close to the line, suggesting that the extremes of model-based inferences are due to the fact that we are selecting subsets, rather than because we are allowing the extra flexibility of admitting different linear combinations of the covariates. As expected from the general theory of Sect. 3.3, judging fit by the F test gives consistently larger values of (31) than when selection is confined to those subsets with $|z| \leq z_\alpha$.

Ferrari and Yang (2015) also consider the set of models selected by the F test at a given level of significance, but instead of discussing the consistency of predictions as we have done here, they are interested in assessing the overall importance of the individual predictors for regression. Their measures are based on looking at incidence and co-appearance of predictors within the subset of the most parsimonious models among those selected. Both approaches illustrate the inadequacy of goodness-of-fit measures such as the F test for choosing predictors, and warn against the uncritical use of conventional methods such as stepwise regression which fail to recognize that equally plausible methods might well give a quite different set of predictors and resulting inferences. The problem of uncertainty in the choice of predictors is particularly challenging in high-dimensional regression, where the number of predictors may be larger than the sample size. The practical importance of such models, in genomics for example, is attracting a large statistical and computational literature on penalized regression methods. See Nan and Yang (2014) for a related discussion of variable selection diagnostics in such models.

5 Comments and conclusions

We have assumed that the significance level α is the same for both confidence intervals and tests of fit. Although this may be the usual convention with $\alpha = 5\%$, the results of Sect. 3 adapt easily to the case of level α for the confidence intervals but some other error rate β for testing model fit. The factor $\sqrt{2}$ appearing in (19) then becomes $m_{\alpha,\beta} = \sqrt{(1 + z_\beta^2/z_\alpha^2)}$. In practice, assessments of fit are often more informal than using a formal goodness-of-fit test, perhaps based on diagnostic plots. Arguably, when looking for specific patterns in such plots it is easy to underestimate sample variability, suggesting that informal model checking may be like taking $\beta > \alpha$. For example if $\alpha = 5\%$ and $\beta = 10\%$, then $m_{\alpha,\beta} = 1.31$ and so the outer confidence limits for ϕ in (19) are wider than CI by 31% instead of by 41%.

The well-known comment by Box (1976), that 'all models are wrong,' continues 'but some models may be useful.' But useful for what? For estimating the unknown parameter ϕ, we might say that M is a useful model if it allows us to narrow down the range of possibilities so that CI_M is a strict subset of CI, the range of values of ϕ that we would otherwise have to entertain in light of the data. But passing a goodness-of-fit test $G(x, M) \le g_\alpha$ is not a sufficient condition for $\mathrm{CI}_M \subset \mathrm{CI}$, as the confidence limits calculated from empirically acceptable models can stray considerably outside the limits of CI. With the usual naive interpretation of confidence intervals, it would only seem sensible for CI_M to include values outside CI if model M was bringing in additional information beyond the information already contained in the data, contradicting the basic frequentist setup of our discussion. The need for such outside assumptions is a common theme of most discussions of model uncertainty. Of the approaches briefly reviewed in Sect. 1, Bayesian model averaging requires (at least) prior probabilities of the different candidate models, frequentist model averaging requires the exact specification of how model weights should depend on the quality of model fit, and marginal assessments of post-selection inference depend on what assumptions are made about the model selection criterion. Examples show that the resulting inferences can depend critically on these outside assumptions.

Model diagnostics are widely used in practice, presumably under the tacit assumption that a good fit means a good model. Such methods are useful for suggesting concise descriptions of the data, but are they useful when the aim is formal inference of the kind considered here? We commonly assume that making modeling assumptions such as using log-linear models for contingency tables or subsets in multiple regression is a good idea because such assumptions give narrower confidence intervals and hence less uncertainty, but ignore the fact that such assumptions carry a bias which varies substantially between different models, even among models that appear to fit the data equally well. The comparison in Sect. 3.3 between $\mathcal{I}$ and $\mathrm{CI}^{(2)}$, the confidence interval we would get under M_0 if we doubled the variance, suggests that the extra uncertainty induced by this bias is of the same order of magnitude as the sampling variability in the data. Ignoring this bias can lead to mutually inconsistent confidence intervals and over-precise inferences.

Our discussion suggests that purely empirical considerations for differentiating between models are not enough. For a model-based analysis to be convincing, we need at least some outside information taking us beyond the observed data. If we have

used a formal method of model selection or model averaging, such outside information is clearly set out, either in the form of the measure to be optimized or in the form of a set of prior distributions. As discussed, the resulting inference can depend very sensitively on what is assumed about these choices. Most users of statistics, however, follow the traditional approach of using a single model and taking the resulting inference at face value, even though the chosen model may be little more than a 'convention' (Hodges 1987). Papers using multiple regression, for example, often assume that only the covariates making a significant contribution to the observed values of the response need to be included, ignoring the fact that other choices of covariates which fit the data equally well can produce sharply differing predictions.

The importance of the model is rarely acknowledged in statistical practice. We need to give much more emphasis to the fact that a conventional model-based confidence interval, or assessment of significance, rests on the assumption that the assumed model is *known to be correct*, and asks wider questions about why this particular model is appropriate in the context of the problem. Merely to show that the model gives an acceptable fit to the data is not enough. The result in Sect. 3.4 that, whatever null hypothesis is being tested, there will always be at least one empirically acceptable model which indicates rejection, shows us that conventional model-based inferences need to be interpreted with considerable caution.

Acknowledgements The authors would like to thank the editors and referees for their very helpful comments on an earlier version of this paper.

References

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*, 791–799.
Claeskens, G., Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press.
Cox, D. R. (1970). *Analysis of binary data*. London: Chapman and Hall/CRC.
Cox, D. R. (1995). Contribution to the discussion of the paper by Draper. *Journal of the Royal Statistical Society, Series B*, *57*, 78.
Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B*, *57*, 45–97.
Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, *109*, 991–1007.
Everitt, B. S. (1977). *The analysis of contingency tables*. London: Chapman and Hall/CRC.
Ferrari, D., Yang, Y. (2015). Confidence sets for model selection by F-testing. *Statistica Sinica*, *25*, 1637–1658.
Hjort, N. L., Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, *98*, 879–899.
Hodges, J. S. (1987). Uncertainty, policy analysis and statistics. *Statistical Science*, *2*, 259–291.
Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417.
Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124.t001.
Langford, J. (2005). Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, *6*, 273–306.
Leeb, H., Potscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, *21*, 21–59.
Miller, A. J. (2002). *Subset selection in regression* (2nd ed.). London: Chapman and Hall/CRC.

Nan, Y., Yang, Y. (2014). Variable selection diagnostic measures for high-dimensional regression. *Journal of Computational and Grahical Statistics*, *23*, 636–656.
Penrose, K., Nelson, A., Fisher, A. (1985). Generalized body composition prediction equation for men using simple measurement techniques (abstract). *Medicine and Science in Sports and Exercise*, *17*, 189.
Potscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, *7*, 163–185.
Royston, P., Sauerbrei, W. (2008). *Multivariate model-building*. Chichester: Wiley.
Simmons, J. P., Nelson, L. D., Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *20*, 1–8.
Wadman, M. (2013). NIH mulls for validating key results. *Nature*, *500*, 14–16.