Annals of the Institute of Statistical Mathematics (2022) 74:615–637
https://doi.org/10.1007/s10463-022-00831-9

INVITED ARTICLE: THIRD AKAIKE MEMORIAL LECTURE

Akaike Memorial Lecture 2020: Some of the challenges of statistical applications

John Copas[1,2]

Received: 8 December 2020 / Revised: 16 July 2021 / Accepted: 19 July 2021 /
Published online: 25 May 2022

Abstract
There has always been a close link between statistical applications and the development of new statistical theory and methods. Even straightforward applications of standard methods can give rise to theoretical challenges leading to new statistical ideas. In my lecture, I will briefly review a few of the statistical developments in my own published papers and describe the applications which gave rise to them. I will then outline some current work on publication bias, one of the outstanding problems in the interpretation of literature reviews, particularly in the medical sciences.

Keywords Shrinkage of predictions · Selectivity bias · Model sensitivity · Publication bias

1 Introduction

Most of the traditional statistical methods that we use were originally developed in response to the needs of particular applications. For example, whilst the great English statistician R. A. Fisher was working at Rothamsted Experimental Station (an agricultural research institute near London), his pioneering papers on the design of experiments, randomization and the analysis of variance provided the statistical foundation for agricultural field trials, now widely used across the experimental sciences. Fisher was a geneticist as well as a statistician, and his important paper on statistical inference in 1920 (exactly 100 years ago this year) established the principles of likelihood and estimation, initially for use in genetics, but now accepted across almost the whole of science.

✉ John Copas
jbc@stats.warwick.ac.uk

[1] University College London, London, UK

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

Of course, papers published in our statistical journals nowadays tend to be much more specialized and technical, whereas the widespread availability of statistical software has meant that research scientists working in application areas can usually carry out their own statistical analyses and publish their work in their own subject-specific journals. However, I think there is still a need for mathematically minded statisticians to collaborate with scientists using statistical methods in their own different application areas, when we will often find that even straightforward applications of standard methods can give rise to statistical challenges which, hopefully, we can then work on in our own research papers.

In my lecture, I would like to illustrate this theme by giving a very brief introduction to three of my own papers published in *JRSSB* (the methodological series of the Journal of the Royal Statistical Society) and describe the applications which gave rise to them. In the final part of my lecture, I will introduce my current research in meta-analysis, again motivated by a rather challenging application. Each of the four sections of this paper is divided into two sub-sections: the application and the resulting paper.

2 Shrinkage of predictions

2.1 Applications

Example 1 During a sabbatical visit to the Naval Postgraduate School in California, I worked on a regression model used by the US Navy for predicting the development costs of new aircraft (Noah et al., 1973). I had access to a vector x of design characteristics for each of 31 previous aircraft and the final costs y of developing those aircraft. To study the difference between retrospective fit and validation fit, I divided the sample into two random subsets, 8 in the 'construction sample' and the remaining 23 in the 'validation sample'. Fitting a linear regression to the construction sample gave the fitted regression

$$\hat{y} = \hat{\alpha} + \hat{\beta}^T x.$$

The 8 points $(y, \hat{y})$ are marked x in Fig 1. As expected, these points are quite close to the line $y = \hat{y}$ shown as the solid line on the plot.

Taking the *same* fitted regression coefficients $\hat{\alpha}$ and $\hat{\beta}$, I then calculated the predicted values of y for the other 23 aircraft. This gave the values $(y, \hat{y})$ for the validation sample marked o on the plot. The solid line now gives a very bad fit to the validation sample, and a straight line fitted to these points, shown as the dotted line, clearly has a very much lower slope than the solid line. Evidently, to get a good fit to the validation sample, we would have to shrink all of the predictions towards the mean. If we judge the worth of a prediction method by its ability to predict new cases, then the original regression is too optimistic, systematically over-estimating its ability to differentiate between cases with lower and higher values of y.

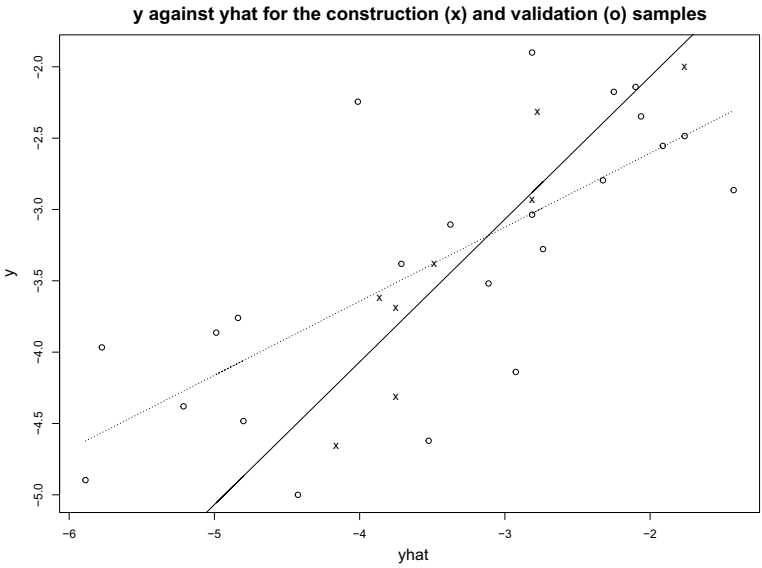

Fig. 1 Example 1. Actual values of y plotted against predicted values $\hat{y}$

Example 2 Shrinkage of predictions is also shown in a second application (Copas and Whiteley, 1976), this time a logistic regression rather than a linear regression. This is one of several criminological applications I worked on through an informal collaboration with the Research Unit of the UK Ministry of Justice. One of the aims of the Research Unit is to devise ways of evaluating the effectiveness of the various criminal penalties which the courts in the UK can give to people who have been found guilty of criminal offences. Depending on the severity of the offence, these can include a prison sentence, a fine, a deferred sentence, supervision under probation, plus a variety of community penalties. Given that randomized controlled trials are not possible in this context, an alternative approach is to rely on prediction studies for specific penalties, using the traditional outcome measure of the presence/absence of further criminal convictions within a fixed period of follow-up.

The second example is a prediction study of convicted criminal offenders whose offending has been related to a specific mental illness (psychopathy) and who have been referred to a specialized psychiatric hospital in London. The binary outcome S/F is defined as the absence/presence of further criminal convictions within a three-year follow-up after leaving hospital, and the input variables x are factors known about these patients at the time of their original conviction. The logistic analysis was based on a sample size of 91 with $p = 6$ covariates.

The simplest way of describing the results of a logistic predictor is to divide the cases into risk groups according to the values of $\hat{z}$, the fitted values of the logits $z =$ logit $P(S|x)$. The plot in Fig. 2 uses five risk groups, based on class intervals of $\hat{z}$ centred on $-2, -1, 0, 1, 2$. For example, the first risk group on the left

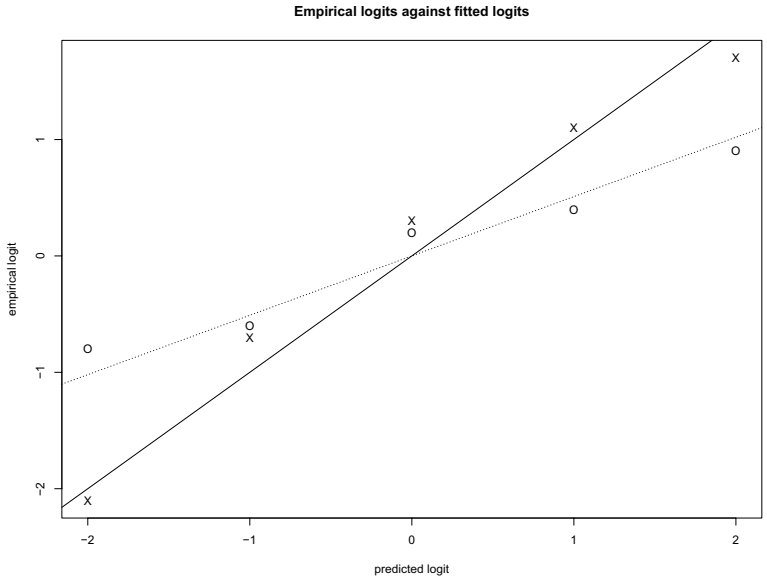

Fig. 2 Example 2. Empirical logits plotted against fitted logits for 5 risk groups

gives a fitted success probability of only 11%, compared to the group on the right with a fitted success probability of 88%. The points marked x in Fig. 2 show, for each risk group, the values of the predicted logit $\hat{z}$ and the empirical logit defined as the sample logit of the observed proportion of successful outcomes within that group. These points are satisfactorily close to the solid diagonal line.

After fitting this logistic model, data from a second sample of patients became available, of similar size to the original sample. To test the validation fit of the fitted model, I used the same fitted logistic regression to predict the values of $\hat{z}$ and the risk groups for the cases in the new sample. The points marked o in Fig. 2 show the corresponding values of the predicted and empirical logits for the new sample. As in the first example, the solid line gives a very poor fit to the validation points. Whilst the predicted probability of success for the risk group on the right is 88%, the empirical logit for this group in the validation sample only gives a success rate of 68%, even lower than the predicted rate for the next lowest risk group. Again, the fitted model is greatly exaggerating the power of the covariates to differentiate between high and low risk cases.

2.2 Copas (1983). Regression, prediction and shrinkage, *JRSSB*, 45, 311–335

The main aim of the paper is to discuss shrinkage of predictions for a number of different regression models and to show how the deterioration in fit between retrospective and prospective samples can be predicted in advance. The simplest model is linear regression, as in the first example, where we have seen that the deterioration in fit corresponds to the difference between the slopes of the solid and dotted lines

in Fig. 1. The linear case is discussed in some detail in Sect. 3 of the paper, with the case of logistic regression discussed in Sect. 8. Here is a simplified version of the arguments to give an outline of the main ideas involved.

Linear Regression. To simplify the notation, assume that both x (the vector of covariates) and y (the outcome variable) are centred about their means. We assume that the observed values of x are multivariate normal with mean zero and variance matrix V, and that $y|x$ follows the usual linear regression model with regression vector β and residual variance σ^2:

$$x \sim N(0, V)\ ,\ y|x \sim N(\beta^T x, \sigma^2)\ .$$

Then for a construction sample of size n, the fitted regression vector is

$$\hat{\beta} \sim N(\beta, n^{-1}\sigma^2 V^{-1})$$

giving the fitted values

$$\hat{y} = \hat{\beta} x.$$

A basic property of least squares shows that the linear regression of y on $\hat{y}$ is simply

$$E(y|\hat{y}) = \hat{y}.$$

For the first example, this is the solid diagonal line shown in Fig. 1.

For the validation sample, suppose we sample independent values of x and y from exactly the same model as before, but now base predictions on the regression vector already fitted from the construction sample. Then, the linear regression of y on $\hat{y}$ for the validation sample is

$$\mathrm{E}\,(y|\hat{y}) = K\hat{y}$$

where

$$K = \frac{\hat{\beta}^T V \beta}{\hat{\beta}^T V \hat{\beta}}\ .$$

The numerator of K depends on the unknown regression vector β, but it can be estimated by noting that

$$\begin{aligned}\mathrm{E}\,(\hat{\beta}^T V \hat{\beta}) &= \beta^T V \beta + \frac{p}{n}\sigma^2 \\ &= \mathrm{E}\,(\hat{\beta}^T V \beta) + \frac{p}{n}\sigma^2\end{aligned}$$

where p is the dimension of the covariate vector x. This suggests the estimate

$$\hat{K} = \frac{n\hat{\beta}^T V \hat{\beta} - p\sigma^2}{n\hat{\beta}^T V \hat{\beta}}$$
$$= 1 - \frac{1}{F}$$

where

$$F = \frac{n\hat{\beta}^T V \hat{\beta}}{p\sigma^2}.$$

This is Fisher's F statistic, the ratio of the regression mean square to the residual mean square in the analysis of variance of the linear regression. The value of F is the usual measure of the strength of evidence for a significant dependence of y on x. A large value of F indicates strong dependence and a value of $\hat{K}$ close to one (little shrinkage). The data in the first example give $F = 3.3$ and $\hat{K} = 0.67$. The line with slope 0.67 has already been shown as the dotted line in Fig. 1, giving a good fit to the validation points marked o on the graph.

Logistic regression. Shrinkage of predictors from logistic regression, as used in the second example, is studied in Sect. 8 of the paper. The fact that logistic regression is a generalized linear model suggests that local approximations to the sampling distribution of maximum likelihood estimates will follow a similar pattern to the linear regression case.

To simplify the notation, suppose that x has first been centred about its mean, so we can assume from now on that x has mean zero with some covariance matrix V. The binary outcome $y = S/F$ is assumed to follow a logistic regression with parameters (α, β), so the model is

$$x \sim N(0, V)\,, \quad \text{logit P}\,(S|x) = \alpha + \beta^T x.$$

For a construction sample of size n, let $\hat{\alpha}$ and $\hat{\beta}$ be the fitted parameters giving the predicted logits

$$\text{logit } \hat{\text{P}}\,(y = S|x) = \hat{\alpha} + \hat{\beta}^T x.$$

To study validation fit, we apply the same prediction model to a new set of independent samples of (y, x) and then fit the *simple* logistic regression model of the true logits on the predicted logits based on the original logistic estimates. Section 8 of the paper shows that the new predicted logits found in this way are approximately

$$\text{logit } \hat{\text{P}}\,(y = S|x) = \hat{\alpha} + \hat{K}\hat{\beta}^T x,$$

where

$$\hat{K} = 1 - \frac{p}{\chi^2}\ ,$$

p is the number of covariates, and χ^2 is the usual deviance statistic for testing the significance of a generalized linear model. For the logistic example here (Example

2), we find that $\hat{K} = 0.51$, corresponding to the dotted line in Fig. 2. The lines with slopes 1 and 0.51 in Fig. 2 give good fits to the construction and validation samples, respectively. Evidently, the ability of the model to discriminate between high risk and low risk cases in the validation sample is *much worse* than would be expected from the fit to the construction sample. The smaller value of K suggests that Example 2 suffers more shrinkage than in Example 1, even though the construction sample size in the second example is considerably larger than in the first example. This may be a general characteristic of binary regression models.

As well as discussing these and other examples, the paper examines connections between the "pre-shrunk" predictors (with the factor K) as suggested here, and fitted models using other approaches to regression. There is a clear similarity with methods of Bayes estimation, which give a positive prior probability to the null hypothesis $\beta = 0$, and also clear links with Stein-type estimates which again imply shrinkage towards $\beta = 0$. An important aspect is the change in dimension from the original regression (with p covariates) to the simple (univariate) regression of y on predicted values $\hat{y}$.

This paper was written almost 30 years ago, and since then, there has been a very considerable expansion of the literature on non-standard regression methods. However, these non-standard methods have been developed by statistical specialists, and until they have been reflected in standard software packages, textbooks and courses, most practitioners will no doubt continue to use the familiar standard regression methods. The approach of the paper is to accept the model as fitted by standard methods and then to study how it can best be used for prediction.

3 Non-random samples

3.1 Application

My interest in this topic started with some joint work with an economist in my university who introduced me to the econometrics literature on "selectivity bias". Nearly all standard statistical methods depend on the assumption that the data we observe are a random sample from the population of interest, even though in practice we routinely use the same methods in observational data analysis. Selectivity bias arises if the randomization assumption is false. The best known example in statistics is missing data—standard methods are only valid if the data are MAR (missing at random), bias arising if the mechanism which determines whether or not data are observed is correlated with the underlying random variables of the assumed model. Another familiar example is in survey analysis when the subjects in the sample are self-selected, each person deciding themselves whether or not to be included in the sample.

We studied a number of interesting examples in economics, but I then came across a clinical trial being conducted in a local hospital which raised very similar issues (Burton and Wells, 1989). The data are shown in Fig. 3. During the 1980's,

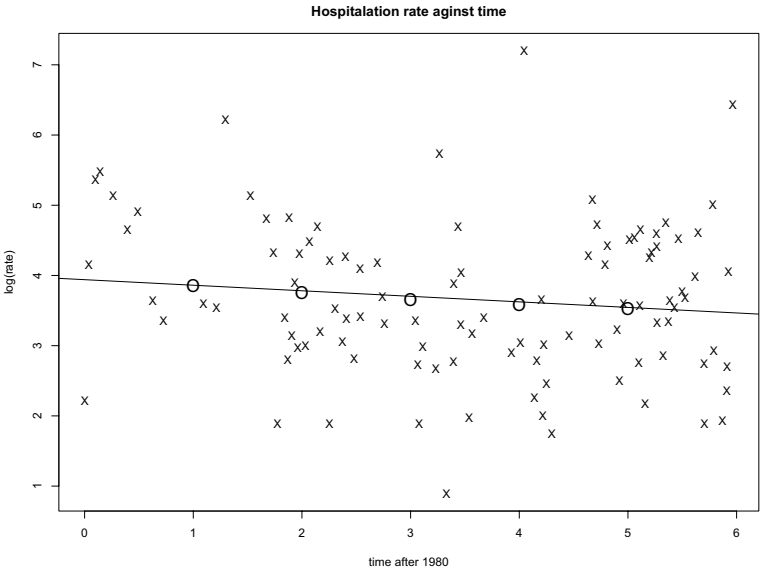

Fig. 3 Hospitalization rate against time

the renal unit of this hospital was testing a new form of kidney dialysis (ambulatory peritoneal dialysis) which was just as clinically effective as the traditional method (haemodialysis) but had a number of practical advantages which might affect the length of time that patients have to stay in hospital. During the trial, patients were allocated to either the new or the old treatment. Figure 3 only includes the data for the new treatment, and the two variables plotted in the graph are x, the date when the patient was entered into the trial (in years from 1980 to 1986), and y, the hospitalization rate defined as the log of the average number of days per year which the patient stayed in hospital.

An unexpected feature of Fig. 3 is the apparent reduction in the value of y as x increases. A linear regression analysis of y on x gives a significant trend, shown as the solid line on the plot. The values of y may well be different between the two treatments, but all the patients featured in the plot have been give the same treatment and so there is no obvious reason why there should be any trend over time. The question of interest in this example is whether this trend could be explained as a consequence of selectivity bias.

3.2 Copas and Li (1997), Inference for non-random samples, *JRSSB*, 59, 55–95.

The example in Fig. 3 is just one of several applications discussed in this paper. One general approach explored in the paper is to develop a local sensitivity analysis for assessing the sensitivity of standard methods of inference to departures from the randomization assumption. The paper defines a scalar sensitivity parameter θ, which reflects the degree of non-randomness in the design of the data ($\theta = 0$ under

the usual randomization assumptions), and then develops local approximations of relevant aspects of inference for small values of θ. However, for the example here, a more direct approach is possible using a special case of the model set out in Sect. 2 of the paper.

3.2.1 Model A

As discussed above, if we make the usual assumption that the data are a random sample, then the linear regression of y on x provides a valid inference, giving the straight line shown in Fig. 3, clear evidence that y does tend to decrease with time. The plot also shows that the number of patients allocated to the new treatment tends to increase over time (more points to the right of the plot). Hospital records of the allocation of patients to the two treatments were used to fit the probit regression model

$$P(S|x) = \Phi(-0.54 + 0.7x) \tag{1}$$

where $S|x$ is the event that a patient was allocated to the new treatment at time x, and Φ is the standard normal distribution function. As regression of y on x is conditioning on x, the fact that the allocation process depends on x is irrelevant as far as the randomization assumption is concerned.

But, further information about this trial showed that this was NOT a randomized controlled clinical trial in the usual sense in medical statistics, raising substantial doubts about the validity of Model A. The allocations of patients to the treatments were not determined by a simple randomizing process but by the doctors who were treating these patients. There may well have been subjective biases influenced by the patients' clinical characteristics which might be correlated with the subsequent values of y. This would clearly undermine the validity of the randomization assumption required by Model A.

3.2.2 Model B

Although the data do suggest that y decreases with x, a reasonable *a priori* assumption would be that y is in fact independent of x—a patient's hospitalization rate should not depend on when the patient happens to have entered the trial. In this case, the model for the values of y should simply be a random sample from, say, a normal distribution with mean β and variance σ^2. The allocation process is modelled by the fitted probit regression model (1).

This leads to a special case of the simultaneous equation model discussed in Sect. 2 of the paper:

$$\begin{aligned} y &= \beta + \sigma\epsilon_1 \\ z &= -0.54 + 0.7x + \epsilon_2 \end{aligned}$$

where (ϵ_1, ϵ_2) is bivariate standard normal with correlation ρ. The first equation specifies the values of y. The second equation specifies a latent selection variable

z: y is observed if and only if $z > 0$. This model is based on papers published in the econometrics literature, in particular Heckman (1976) and Heckman (1979).

The second equation gives

$$\mathrm{P}\,(S|x) = \mathrm{P}\,(z > 0|x) = \Phi(-0.54 + 0.7x)$$

which is exactly the same as Eq. (1) above.

Using standard properties of the bivariate normal distribution, we get

$$\mathrm{E}\,(y|x, S) = \mathrm{E}\,(y|x, z > 0) = \beta + \rho\sigma\Lambda(-0.54 + 0.7x) \tag{2}$$

where $\Lambda(u) = \phi(u)/\Phi(u)$ is Mill's Ratio, the ratio of the standard normal density to the standard normal distribution function. Similarly, the model also gives the variance

$$\mathrm{Var}\,(y|x, S) = \sigma^2[1 - \rho^2\Lambda(-0.54 + 0.7x)\{-0.54 + 0.7x + \Lambda(-0.54 + 0.7x)\}]. \tag{3}$$

Equations (2) and (3) specify a nonlinear regression model for the observed values of y and x, which can be fitted by weighted least squares in the usual way, giving estimates of the remaining parameters β, σ and ρ. Substituting these estimates into (2) gives the fitted regression model under Model B. Evaluating this function for $x = (1, 2, \cdots, 5)$ gives the points marked O in Fig. 3. *These values are almost indistinguishable from the linear regression line fitted under Model A*. The Mills ratio Λ is a convex function, but evidently the values x used here fall into the range when it is close to linearity.

These data are a good example of the importance of randomization assumptions in statistical analysis. We have two different models, A and B. Model A assumes the data are random with selection function (1), an ignorable model in the sense that allocation depends on x but not on y. Model B, on the other hand, implies the allocation model

$$\mathrm{P}\,(S|x, y) = \mathrm{P}\,(z > 0|x, y) = \Phi\left\{\frac{-0.54 + 0.7x + \rho(y - \beta)/\sigma}{\sqrt{(1 - \rho^2)}}\right\} \tag{4}$$

which, when $\rho \neq 0$, is non-ignorable since the allocation explicitly depends on both x *and* the dependent variable y. However, Fig. 3 shows that the *fitted* regression of y on x is almost exactly the same, both models showing a clear observed decline in y over x. But the *interpretation* of the models is quite different: model A allows for a structural dependence of y on x, model B asserts that the there is only an *apparent* dependence between the two variables, the observed dependence being simply a consequence of the sample selectivity specified in Eq. (4). Under model A, the fitted slope of y on x is an unbiased estimate of the true slope. Under model B, the true slope is known to be zero, the fitted slope being entirely caused by selectivity bias.

Just by looking at the data (the scatter of points in Fig. 3), it is impossible to know which of these two models is more likely to be correct and therefore impossible to have a clear interpretation of the results of the analysis. Model choice must depend on information outside of the data; in this case on the way, the hospital has

allocated the patients to the treatments. More generally, for a statistical analysis to be convincing, its assumptions have to reflect the scientific context of the experiment, which gave rise to the data, and not just chosen on the basis of the data alone.

A more general version of Model B could allow y in the first equation to also involve covariates. For example, if β in the first equation is replaced by $\beta_0 + \beta_1 x$, then the extended model also includes Model A as the special case when $\rho = 0$. Fitting the extended model with ρ as an unknown parameter, however, depends on the ability of the weighted least squares algorithm to differentiate between two covariates, x and the nonlinear function of x involved in Eq. (2). The near linearity of the points marked O in Fig. 3, however, means that these two covariates will be almost collinear, and so the weighted least squares algorithm will be very unstable. This problem, and more general versions of it, have been much discussed in the econometrics literature, leading to the suggestion that models of this type (the Heckman method) should only be used when the y-equation and the z-equation are based on *different* covariates.

4 Model sensitivity

4.1 Application

This application arose from discussions within the Methodology Committee of the Office of National Statistics (ONS). The ONS is the central statistics department of the UK government, and the task of the committee is to discuss and comment on the statistical methods being used by the government statisticians. It is another very rich source of interesting statistical problems.

One of the responsibilities of the ONS is to compile regular statistical reports on the incomes (or wages) received by the working population of the UK. An obvious question is, given data on incomes, how do we estimate the mean income? This is not a trivial question, because income distributions tend to be highly skewed and so sample averages can be very sensitive to outliers. My suggestion to the committee was that, instead of using one of the methods suggested in the literature for down-weighting the effect of outliers, it might be better to fit a parametric model which allows for skewness and then to use the fitted parameters of the model to estimate the expectation.

The simplest possibility is to assume that income has a log-normal distribution, and so

$$X = \log(\text{ income }) \sim N\left(\theta, \sigma^2\right).$$

Then the mean income is

$$\mathrm{E}\left\{\exp(X)\right\} = \exp\left(\theta + \frac{1}{2}\sigma^2\right),$$

and hence the maximum likelihood estimate of

$$\phi = \log\{\ \mathrm{E}\ (\exp(X)\},$$

the log of the mean income, is simply

$$\hat{\phi} = \bar{x} + \frac{1}{2}s^2,$$

where $\bar{x}$ and s^2 are, respectively, the sample mean and sample variance of the data values of log(income).

The log-normal model gives a good fit to the observed sample of incomes, but we know that a good fit to the data does not necessarily mean that the model is known to be correct. The log-normal model may be just one of a family of alternative models, G say, which also give acceptable fits to the same data. The fact that the data will give us very little reason to discriminate between the different models within this set suggests that when we assess the uncertainty in $\hat{\phi}$ we should also include the uncertainty arising from the choice of the different models within G. So the challenge arising from this simple example is how to develop a sensitivity analysis which allows for uncertainty both within and between the class of well-fitting models.

4.2 Copas and Eguchi (2010) , Likelihood for statistically equivalent models. *JRSSB*, 72, 193–217

This in one of several joint papers published with Professor Shinto Eguchi of the ISM, giving a general likelihood theory for problems similar to that in the above example. The basic theory is put forward in Sects. 2.1–3 of the paper, illustrated by the log-normal incomes example in Sect. 2.4. The term "statistically equivalent models" in the title of the paper refers to the set of alternative models like G in the example, models which could also have been chosen as comparable alternatives to the working model which was originally fitted to the data. All models considered in the paper are assumed to satisfy the usual regularity assumptions needed for first order likelihood asymptotics.

4.2.1 Basic set-up

Suppose that the working model, f, for an observed sample $x_1, x_2, \ldots, x_n$ is

$$f\ :\ X \sim f(x, \theta)\ .$$

We assume that model f has been chosen because it gives a good fit to the data.

However, there will also be many other well-fitting models which could also have been chosen for the analysis. For any such alternative model, g say, we assume the decomposition

$$g : X \sim g(x) \propto \exp\{\epsilon u(x,\theta)\} f(x,\theta) \ .$$

If $\epsilon = 0$ then $g = f$, so we can think of ϵ as the *distance* between g and f. The fact that both g and f give good fits to the same data suggests that ϵ will be small. Section 2.1 of the paper shows that, under reasonable regularity assumptions,

$$\epsilon = O\left(n^{-\frac{1}{2}}\right) \ .$$

Similarly, we can think of the function $u(x,\theta)$ as the *direction* of the displacement between g and f. With no loss of generality, we can assume that $u(x,\theta)$ is standardized so that

$$\mathrm{E}_f u(X,\theta) = 0 \quad \text{and} \quad \mathrm{Var}_f u(X,\theta) = 1 \ .$$

Since

$$u(x_i,\theta) \propto \log\{g(x_i)/f(x_i)\} \ ,$$

the asymptotic log-likelihood ratio test of g against f accepts g at level α if

$$|S_u| \le z_\alpha \ ,$$

where

$$S_u = n^{-\frac{1}{2}} \sum_{i=1}^{n} u\left(x_i,\theta\right) \ ,$$

and z_α is the appropriate percentage point of the standard normal distribution. We define G as the set of models g which are accepted by this likelihood ratio test. As only local departures from f are involved, the paper shows that the parameter θ appearing in these expressions can safely be replaced by its maximum likelihood estimate under model f.

As the parameter θ only has meaning within the model f, we need to define the *parameter of interest* as a characteristic of the problem rather than of the model which happens to have been selected. The paper does this by defining, for any model g, the parameter of interest, ϕ, to be the solution of the estimating equation

$$\phi : \ \mathrm{E}_g a(X,\phi) = 0 \ .$$

For example, in the mean incomes problem, the estimating function $a(x,\phi)$ is

$$a(x,\phi) = 1 - \exp(x - \phi) \ .$$

Under model f, this implies that $\phi = \log\{E_f \exp(X)\}$ as before.

4.2.2 Likelihood functions

If $\hat{\phi}_f$ and $\hat{\sigma}_f^2$ are respectively the maximum likelihood estimates of ϕ and its variance under model f, define the pivital function

$$\omega(\phi) = \frac{n^{\frac{1}{2}}\left(\phi - \hat{\phi}_f\right)}{\hat{\sigma}_f} .$$

The asymptotic log-likelihood function for ϕ under f is then

$$L_f(\phi) = -\frac{1}{2}\{\omega(\phi)\}^2 .$$

Let $L_g(\phi)$ be the corresponding asymptotic log-likelihood function based on model g. These functions, for $g \in G$, define a family of likelihoods which, depending on the direction function u, may be displaced to the right or to the left of $L_f(\phi)$. If we are interested in the range of values of ϕ which could be considered plausible under at least one model $g \in G$, then it is the envelope of this family of likelihoods,

$$L_{ENV}(\phi) = \sup_{g \in G} L_g(\phi)$$

which we need to consider. Section 2.3 of the paper shows that

$$L_{ENV}(\phi) = -\frac{1}{2}\left[\max\{\rho|\omega(\phi)| - (1-\rho^2)^{\frac{1}{2}} z_\alpha, 0\}\right]^2 ,$$

where

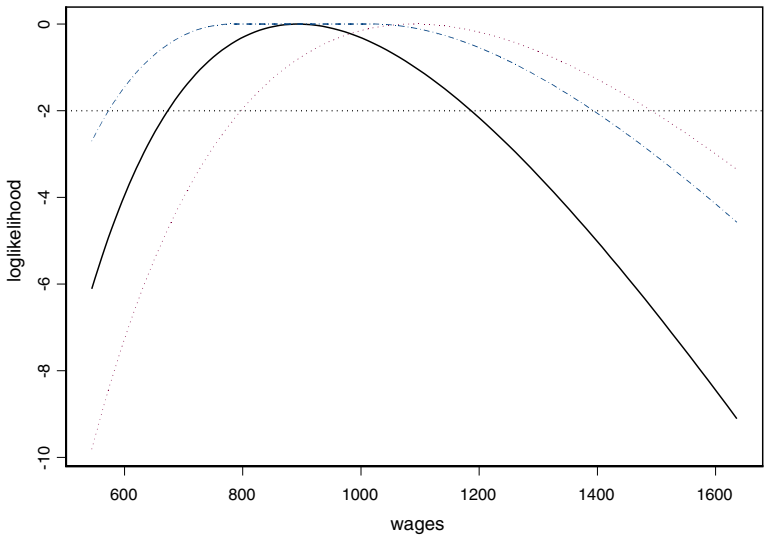

Fig. 4 Log likelihoods for mean income

$$\rho = \text{corr}_f\{a(X, \phi), s(X, \theta)\}$$

and $s(x, \theta)$ is the score function of the model f. The parameter ρ plays a crucial role in this discussion. If $\rho = 1$, $L_{ENV} \equiv L_f$ and so g is essentially just a re-parameterization of model f. As ρ decreases from 1, L_{ENV} becomes increasingly dispersed until, when $\rho = 0$, the envelope likelihood is identically zero for all ϕ, indicating that ϕ is then unidentifiable (no inference is possible).

Example The data used in the ONS study of incomes mentioned earlier are confidential and so I cannot illustrate the above theory on the actual data used, but I can show these likelihood functions for a smaller simulated data set having very similar characteristics to actual incomes in the UK in 2002. The sample size is $n = 100$, much smaller than the actual data set used in the application. The data are values of $X = \log$ income, and the parameter of interest is $\phi = \log\{\text{ E } \exp(X)\}$ as before. The working model f assumes that X is normally distributed.

The solid line shown in Fig. 4 is $L_f(\phi)$ plotted against $\exp(\phi)$, the values of the actual mean income or wages (transforming the horizontal axis from ϕ to $\exp(\phi)$ accounts for the asymmetry of this likelihood function). $L_{ENV}(\phi)$ is plotted against income to give the dashed line in the plot.

A standard asymptotic property of regular univariate log-likelihood functions $L(\theta)$ is that the approximate 95% confidence limits for θ are given by the two solutions of

$$L(\theta) = \sup_{\theta} L(\theta) - 2.$$

Thus, if the maximum of $L(\theta)$ is (arbitrarily) set to zero, the confidence limits are given by the intersections of L with the horizontal line at $L = -2$.

Applying this to L_f in Fig. 4 gives the confidence limits for mean income per week as (£ 680, £ 1200), an interval of width £ 520. Applying this to L_{ENV} gives intersections at (£ 560, £ 1400). We do not know the confidence intervals for individual functions g, but we do know that, over the set $g \in G$, the lower confidence limit can be as low as £ 560, and the upper confidence limit can be as high as £ 1400, a difference of £ 840. If we interpret the width of a confidence interval as a crude measure of uncertainty, then a conservative estimate of the effect of model uncertainty in this example is that it may have increased the overall uncertainty in the estimate of mean income by about 60%. Of course the sample size in the ONS analysis is much larger than in the example here, and so the width of confidence intervals will be much narrower, although the relative uncertainties of different methods are likely to be quite similar.

Figure 4 also shows another log-likelihood function given by the dotted line, L_{SP}, described in Sect. 2.3 of the paper as the semi-parametric likelihood. This is centred on the nonparametric estimate of ϕ given by

$$n^{-1} \sum a\left(x_i, \tilde{\phi}\right) = 0\,,$$

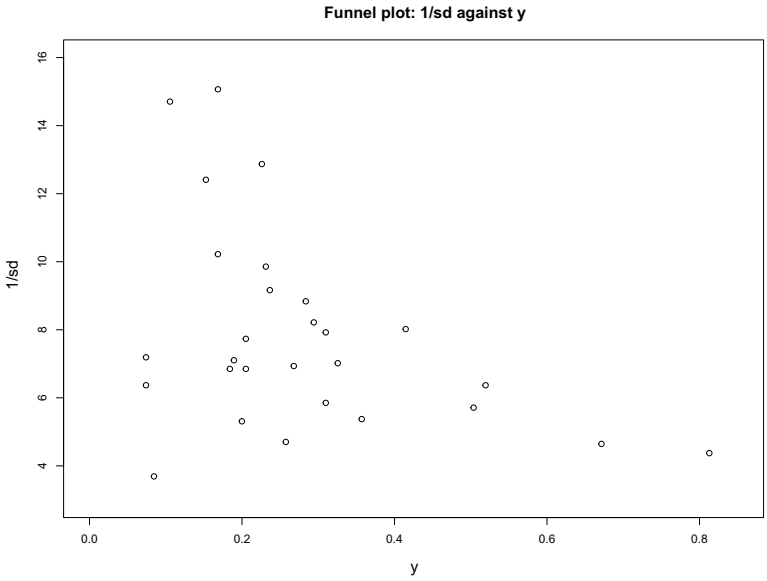

Fig. 5 Funnel plot of the criminological review

which for $a = 1 - \exp(x - \phi)$ is simply the sample mean of the observed incomes. The shape of L_{SP} is determined by assuming that $\tilde{\phi}$ is normally distributed with mean ϕ and variance taken to be the same as its variance under model f.

The sample values of income used in this example have one rather large and influential observation, which has a marked effect on the sample average of the actual incomes but relatively little effect on the sample average of the log incomes. As expected, removing this outlying observation from the sample has a much more marked effect on L_{SP} than on L_f or its neighbouring likelihoods L_g with $g \in G$.

5 Publication bias

5.1 Application

The second example in Sect. 2.1 of this paper arose from my collaborative work with the UK Ministry of Justice. The application in this section, shown in Fig. 5, also arose from this collaboration. Figure 5 is a funnel plot of a meta-analysis which appeared in the Journal *Criminology* in 1990, probably the first published systematic review in this area (Andrews et al., 1990). It was received with considerable interest at the time, since it was one of only very few research studies in criminology which claimed to provide clear evidence that a newly developed policy for the supervision of juvenile offenders actually 'works' in the sense of reducing subsequent offending. I was asked to comment on the quality of the statistical methods being used in this paper. Although standard methods

of meta-analysis had been followed, it was only when I plotted the data, resulting in Fig. 5, that there appeared to be a substantial problem of publication bias, indicated by the very marked skewness of this plot. This is likely to have severely exaggerated the main conclusion of the review.

The review studied the results of 29 separate clinical trials, each comparing the effectiveness of a new form of supervision of convicted juvenile offenders with the traditional form, measuring success as the absence of any further criminal convictions within a fixed period of follow-up. Each study results in a 2×2 table of the numbers of successes and failures for each of the two treatments, and hence a value of the traditional χ^2 statistic. The meta-analysis was based on a fixed effects model, defining the treatment effect estimate for the ith study, with its asymptotic sampling distribution, as

$$y_i = \pm\sqrt{\chi_i^2/n_i} \sim N(\theta, n_i^{-1}) \,,$$

where n_i is the study sample size and $\pm$ indicates whether the observed success rate for the new treatment was higher or lower than the rate observed for the controls. This way of defining the treatment effect has traditionally been used in psychology and the social sciences, although most recent research in the medical sciences has almost always defined y as the log odds ratio or the log relative risk, with the variance taken to be the value of its usual estimate.

The funnel plot in Fig. 5 is simply the plot of $1/sd_i = \sqrt{n_i}$ on the vertical axis against y_i on the horizontal axis. If the above fixed effects model is correct, then these points should look like a funnel, with values of y_i clustered about a common value of θ with spread increasing as we move from the top to the bottom of the plot. However, this is clearly not the case in Fig. 5—evidently the smaller studies (with smaller sample sizes n_i) are skewed to the right, tending to give larger values of y_i than would be expected from the more accurate estimates nearer the top of the plot. This is the "small study effect", usually taken as a sign of publication bias.

The challenge of this application is to find a relatively simple model with interpretable parameters which can describe skewness observed in a funnel plot and so suggest a sensitivity method for assessing publication bias.

5.2 Current work on publication bias

My discussion of the Andrews review, and of similar examples in the medical literature, are reflected in a number of subsequent statistical papers. My main concern in these papers has been to develop a sensitivity analysis which indicates the likely effect of publication bias on the main results of the analysis. The method proposed in Copas and Shi (2000), sometimes called the 'Copas method', has been implemented in R software and used in a number of systematic reviews. The method is based on an adapted version of the econometric model used earlier in Sect. 3.2 above.

However, this method is not without its problems. Although the non-randomness correlation parameter ρ is well-defined mathematically, it has no clear interpretation in terms of identifiable aspects of the problem, and so the dependence of results

on ρ can be difficult to interpret. The second problem is that in some examples, the numerical algorithm can fail to converge. Current work aims to overcome these problems by developing a simpler model with interpretable parameters which can be used without the need for complicated numerical algorithms.

Section 3 of Copas (2013) developed a general theory of selection functions for publication bias, based on the assumptions that the set of studies which have been published (and hence available for inclusion in a systematic review) are a non-random selection from the larger population of all the studies which have been carried out in the area of interest. A selection function assumes that relevant outcomes of the wider population of studies are given by values of a random variable x, and that studies with outcome x are selected with probability

$$\mathrm{P}\,(\text{ select }|x) = a(x)$$

for some probability function $a(x)$. The aim is to make an inference about the distribution of x across the wider population of studies, given only the data on values of x within the smaller population of selected studies. Clearly, this inference will depend critically on the choices of x and $a(x)$. Usually, little or no information is available about the unpublished studies, in which case the choices of x and $a(x)$ can only depend on knowledge of the context of the studies and on the data observed in the studies which have been published, usually summarized by a funnel plot as in Fig. 5. Assessing the effect of publication bias is one of the most difficult problems in meta-analysis, and in practice, it is usually ignored altogether. This is equivalent to assuming that publication is a purely random process, with $a(x) \equiv p$ for some arbitrary constant p. The tacit assumption that journal editors decide to publish submitted papers purely at random seems entirely implausible.

Taking the criminological review as an example of clinical trials comparing two treatments, suppose that, for each trial, y is the estimated treatment effect with variance σ^2, satisfying the fixed effects model

$$y \sim N\left(\theta, \sigma^2\right) \ .$$

Assume that the sample size in each trial is sufficiently large that we can ignore the sampling errors of the within-trial variance estimates. The usual within-trial significance test of $H_0 : \theta = 0$ is to refer $t = y/\sigma$ to the standard normal distribution. For any value of θ,

$$t = \frac{y}{\sigma} \sim N(\phi, 1)$$

where

$$\phi = \frac{\theta}{\sigma} \ .$$

There is substantial evidence, at least in the medical literature, that studies reporting a significant difference between the two treatments are more likely to be

published than studies reporting non-significant results (Dwan et al., 2013). If a 2-tail test is appropriate, and S is the event of selection (publication), then this suggests that P $(S|t)$ should be an increasing function of $|t|$. A simple example of such a selection model is the exponential model

$$\mathrm{P}\,(S|t) = 1 - \exp\{-(\alpha + \beta t^2)\} \tag{5}$$

with two selection parameters $\alpha \geq 0$ and $\beta \geq 0$. If $\beta = 0$ then selection is independent of t and hence selection is a purely random process (and hence ignorable). The size of β is the principal determinant of publication bias, with α controlling the marginal probability of selection (the overall proportion of papers which are published).

Equation (5) gives

$$\begin{aligned}\mathrm{P}\,(t \cap S|\sigma) &= \mathrm{P}\,(t|\sigma)\,\mathrm{P}\,(S|t,\sigma)\\ &= \frac{1}{\sqrt{(2\pi)}}\left[\exp\left\{-\frac{1}{2}(t-\phi)^2\right\} - \exp\left\{-\frac{1}{2}(t-\phi)^2 - \alpha - \beta t^2\right\}\right].\end{aligned}$$

Completing the square in this last expression allows us to write it as a linear combination of two normal density functions $f_1(t)$ and $f_2(t)$, giving

$$\mathrm{P}\,(t \cap S|\sigma) = f_1(t) - \left\{(1+2\beta)^{-\frac{1}{2}} \exp\left(-\alpha - \frac{\beta}{1+2\beta}\phi^2\right)\right\} f_2(t) \tag{6}$$

with

$$f_1(t) \sim N(\phi, 1) \quad \text{and} \quad f_2(t) \sim N\left(\frac{\phi}{1+2\beta}, \frac{1}{1+2\beta}\right).$$

Integrating (6) over t gives

$$\mathrm{P}\,(S|\sigma) = 1 - \frac{1}{\sqrt{1+2\beta}} \exp\left\{-\alpha - \frac{\beta}{1+2\beta}\phi^2\right\}. \tag{7}$$

Multiplying (6) by t and then integrating over t gives the conditional mean

$$\mathrm{E}\,(t|S,\sigma) = \phi \frac{2\beta + \mathrm{P}\,(S|\sigma)}{(1+2\beta)\,\mathrm{P}\,(S|\sigma)}. \tag{8}$$

Similarly, we can get the variance and other properties of the distribution of t for the selected studies.

Making these assumptions, the log-likelihood function $L(\theta, \alpha, \beta)$ for any observed meta-analysis follows immediately from Eqs. (5) and (7). If t_i and ϕ_i are the values of t and ϕ for the ith study, and S_i is the event that the ith study is selected, then, omitting irrelevant additive constants,

$$L(\theta, \alpha, \beta) = \sum_{i|S_i} \left\{-\frac{1}{2\phi_i^2}(t_i - \theta)^2 + \log\ \mathrm{P}\,(S_i|t_i) - \log\ \mathrm{P}\,(S_i|\sigma_i)\right\}.$$

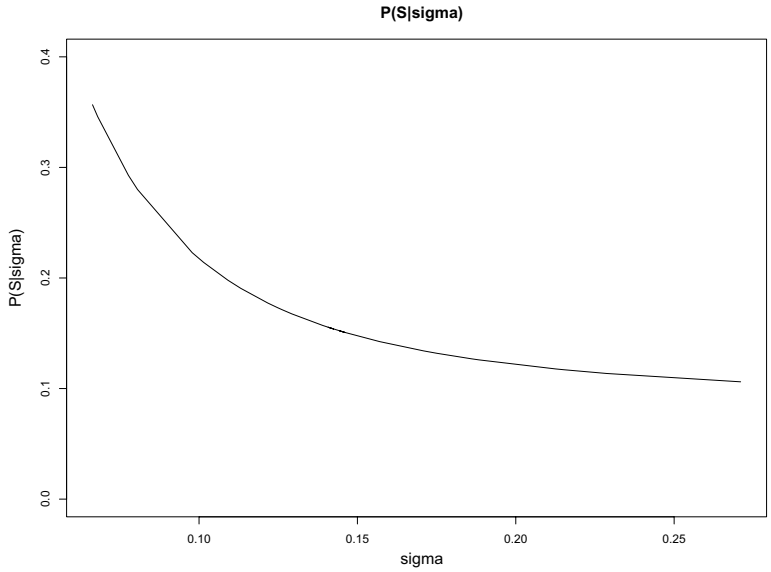

Fig. 6 P $(S|\sigma)$ as a function of σ

A simple numerical search of the values of this log-likelihood function for the data in Fig. 5 shows that, for this example, the maximum likelihood estimates of (θ, α, β) are approximately

$$\hat{\theta} = 0.16, \quad \hat{\alpha} = 0.00, \quad \hat{\beta} = 0.01 \quad .$$

Here are four ways of illustrating the fit of this model to the criminological data in Fig. 5 :

(a) *Severity of Selection*

Figure 6 shows a plot of the estimates of (7), the probability of selection for values of σ within the range observed in the meta-analysis. Evidently, the largest study (with smallest σ) is three times more likely to be published than the smallest study (with largest σ). The dependence of selection on the size of t is much stronger than in the naive "missing at random" model tacitly assumed in most applications of meta-analysis.

(b) *The Fitted Funnel Plot*

Writing $y = t\sigma$ and $\theta = \phi\sigma$, Eq. (8) gives

$$\mathrm{E}\,(y|S, \sigma) = \sigma\ \mathrm{E}\,(t|S, \sigma) = k_\sigma \theta \quad (9) \tag{9}$$

where

$$k_\sigma = \frac{2\beta + \ \mathrm{P}\,(S|\sigma)}{(1 + 2\beta)\ \mathrm{P}\,(S|\sigma)} \ .$$

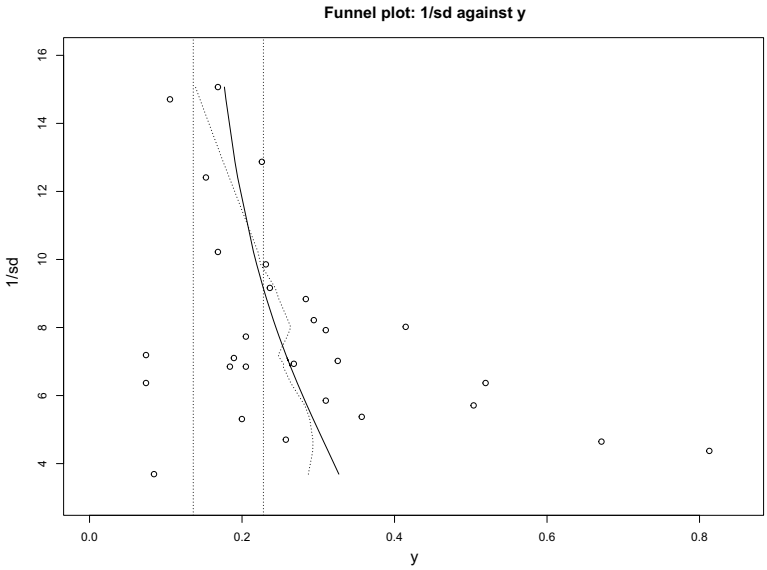

Fig. 7 The fitted funnel plot

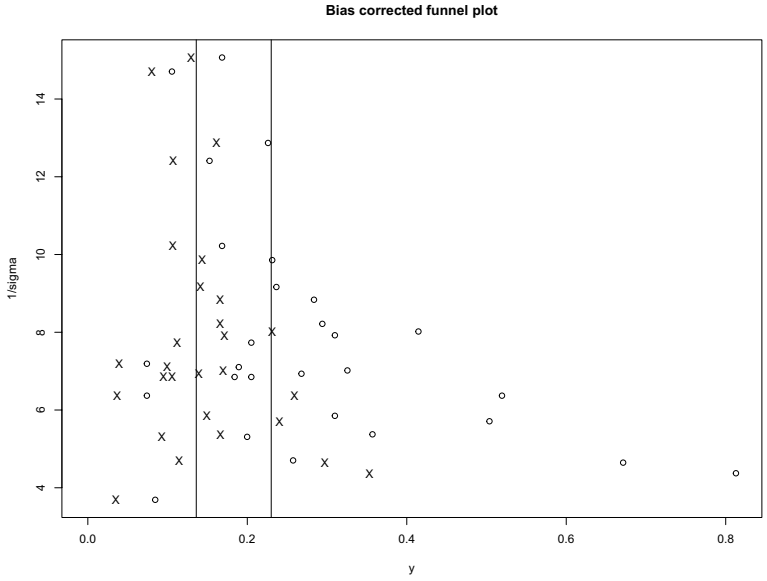

Fig. 8 The bias-corrected funnel plot

As P $(S|\sigma)$ decreases with σ (see Fig. 6), the factor k_σ is an increasing function of σ. This means that the fitted value E $(y|S, \sigma)$ increases from θ, for very large studies

(at the top of the funnel plot), to progressively larger values towards the bottom of the plot. This is seen in Fig. 7, which shows values of $\mathrm{E}(y|S,\sigma)$ as the solid line superimposed on the funnel plot of Fig. 5.

Also shown in Fig. 7 is a LOWESS scatter plot smoother of the observed values y_i against $1/\sigma_i$ (the dotted line). The fact that the solid line is quite close to this crude nonparametric estimate of the regression of y on $1/\sigma$ suggests that the fitted values given by the model give a good estimate of the skewness of the funnel plot.

(c) *The Bias-corrected Funnel Plot*
Equation (5) shows that

$$\tilde{y}_i = y_i/k_\sigma$$

is an unbiased estimate of θ for all values of σ. The plot of $1/\sigma_i$ against $\tilde{y}_i$ is therefore a bias-corrected version of the funnel plot. These points are added to the original funnel plot of the criminological review in Fig. 8—the original points are shown as o (as before), and the bias-corrected points are shown as x. The shape of the new points are as one expects of a funnel plot, successfully removing the skewness of the original plot.

(d) *Estimates of* θ
The usual estimate of θ in fixed effects meta-analysis is the weighted average

$$\hat{\theta} = \left\{\sum \sigma_i^{-2}\right\}^{-1} \sum \left(y_i \sigma_i^{-2}\right) \ .$$

For the criminological data, this gives

$$\hat{\theta}_1 = 0.228 \ .$$

But for the bias corrected estimate, we replace y_i by $\tilde{y}_i$ to give

$$\hat{\theta}_2 = 0.163 \ .$$

As expected, this estimate is close to the maximum likelihood estimate of θ found earlier.

These two estimates are shown as the vertical lines in Figs. 7 and 8. Evidently, for these data, publication bias has resulted in the conventional analysis over-estimating the treatment effect by about 40%. In most applications of statistics, a bias of this magnitude would be thought of as a very major problem.

Although the model proposed here seems to give a good description of features of the funnel plot in Fig. 5, further research is clearly needed on the theoretical properties of the model and its suitability for use in other examples. An extension of the model to allow for random effects is clearly needed. Although a variance component could be estimated in the usual way, this would ignore the effect of the selection model assumed here. It may be possible to add a variance component parameter directly into the likelihood function. A major practical problem is that most systematic reviews have fewer studies than in the example used here, which would make

it much more difficult to estimate the selection parameters α and β. The review by Dwan et al. (2013) referred to earlier provides some empirical estimates of the proportion of studies reporting significant treatment effects in published studies, as well as in unpublished studies. The differences between these estimates across a number of different medical areas can provide estimates of α and β which might provide a prior distribution for a Bayesian analysis. Hopefully, further research on these and other aspects will be reported in a future publication.

References

Andrews, D. A., Zinger, I., Hope, R., Bonta, J., Gendreau, P., Cullin, F. T. (1990). Does correctional treatment work?: A clinically relevant and psychologically informed meta-analysis. *Criminology, 28,* 369–429.

Burton, P. R., Wells, J. (1989). A selection adjusted comparison of hospitalization on continuous ambulatory peritoneal dialysis and haemodialysis. *Journal of Clinical Epidemiology, 42*, 531–535.

Copas, J. B. (1983). Regression, prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society, B, 45*, 311–354.

Copas, J. B. (2013). A likelihood-based sensitivity analysis for publication bias in meta-analysis. *Applied Statistics, 62*, 47–66.

Copas, J. B., Eguchi, S. (2010). Likelihood for statistically equivalent models. *Journal of the Royal Statistical Society, B, 72*, 193–217.

Copas, J. B., Li, H. G. (1997). Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society, B, 59*, 55–95.

Copas, J. B., Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics, 1*, 247–262.

Copas, J. B., Whiteley, J. S. (1976). Predicting success in the treatment of psychopaths. *British Journal of Psychiatry, 129*, 388–392.

Dwan, K., Gamble, C., Williamson, P. R., Kirkham, J. J. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. *PLOS ONE, 8*(7), e66844.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement, 5*, 475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*, 153–161.

Noah, J. W., Daniels, J. M., Day, C. F., Eskew, H. L. (1973). Estimating aircraft acquisition costs by parametric methods. *United States Navy*, FR-103-USN.