Meta-Analysis, Funnel Plots and Sensitivity Analysis

John Copas[1] and Jian Qing Shi
University of Warwick, UK

SUMMARY

Publication bias is a major problem, perhaps *the* major problem, in meta-analysis (or systematic reviews). Small studies are more likely to be published if their results are "significant" than if their results are negative or inconclusive, and so the studies available for review are biased in favour of those with positive outcomes. Correcting for this bias is not possible without making untestable assumptions. In this paper, a sensitivity analysis is suggested which is based on fitting a model to the funnel plot. Some examples are discussed.

Keywords: FUNNEL PLOTS; META-ANALYSIS; SELECTIVITY BIAS; SENSITIVITY ANALYSIS.

[1] *Address for correspondence*: Department of Statistics, University of Warwick, Coventry CV4 7AL, UK: email: jbc@stats.warwick.ac.uk

1. Introduction

Because of a rapidly expanding volume of scientific research, meta-analysis is becoming increasingly important as a way of collecting and synthesizing results from individual studies. Examples abound in medical research — particularly in systematic reviews of clinical trials and epidemiological studies. Many of these studies are reporting small experimental or epidemiological investigations which, individually, may fail to come to any very firm conclusion about the treatments or risk factors being compared, but collectively may suggest a clear overall result.

Two important statistical difficulties in combining such research studies are *heterogeneity* and *publication bias*. As our initial example, we consider the problem of assessing the relationship between passive smoking and lung cancer, a topic of much current debate (see, *e.g.*, Poswillo *et al.*, 1998, Hackshaw *et al.*, 1997; Givens *et al.*, 1997). Hackshaw's paper (Hackshaw *et al.*, 1997) reviewed 37 published epidemiological studies of the risk of lung cancer in female non-smokers whose spouses/partners did or did not smoke. Each of these studies reported an estimate of the relative risk (odds ratio) and a 95% confidence interval. Most of the 37 studies found an increased risk in the exposed group, but a few came to the opposite conclusion. The points in Figure 1 plot the log odds ratio y_i against the standard error s_i, the so called "funnel plot" for these 37 studies. We have taken the values of s_i as simply one quarter of the widths of the log-transformed confidence intervals.

If μ is the overall log odds ratio, then the crude (fixed effects) estimate of μ is the weighted (pooled) average of the log odds ratios for all 37 studies, which comes to 0.19 (relative risk is 1.21). If there is no heterogeneity, we should expect a majority of studies (about 70% of them) to report values of y in the range $(\mu - s, \mu + s)$ (the dotted lines in Figure 1). The variation of the points for the large studies in Figure 1 is larger than expected; there is clear evidence of heterogeneity. Hackshaw *et al.* (1997) recognised this by using a random effects rather than a fixed effects analysis — this has the effect of giving more weight to the smaller studies, raising the log-odds ratio from 0.19 to 0.22.

Clearly the results from the smaller studies will be more widely spread around the average effect because of their larger standard errors, but they should also be symmetrically placed about the line $y = \mu$, to form the shape of a "funnel". Figure 1 however shows a clear tendency for the smaller studies to give more positive results than the larger studies — this is a sign of publication bias. The suspicion is that there are other small studies which have been carried out but which have not been published, and that those selected for review are biased in favour of those with positive outcomes. In practice, publication bias is a very common and very serious problem. Sutton *et al.* (1999), for example, assessed 48 meta-analyses and found that about half showed signs of potential publication bias similar to that evident in our example here.

Several methods for "correcting" publication bias have been proposed in the literature. These include selection models using weighting functions (*e.g.* Hedges, 1984, 1992; Iyengar and Greenhouse, 1988; Dear and Begg, 1992; Silliman, 1997a), Bayesian methods (e.g. Givens *et al.*, 1997; Silliman, 1997b) and the "Trim and Fill" method (Taylor and Tweedie, 1998a, b). Necessarily such methods make unverifiable assumptions, for example the Bayesian approach by assuming a prior distribution on the number of unpublished studies, and the trim and fill method by making strong symmetry assumptions. An alternative more cautious approach is *sensitivity analysis* – we draw conclusions from the meta-analysis under a variety of plausible possibilities for the extent of publication bias, and assess how different these conclusions are from one another and from the results of standard approaches. A rather general methodology for doing this will be proposed in Section 2 following the earlier work by Copas and Li (1997) and Copas (1999). A detailed procedure for sensitivity analysis is presented in Section 3, illustrated on the example of passive smoking and lung cancer. A non-technical account of our reanalysis of these data is in Copas and Shi (2000), where we conclude that the published estimate of the increased risk of lung cancer associated with environmental tobacco smoke may be overestimated. Some more examples are presented in Section 4, and some final comments included in Section 5.

The aim of a sensitivity analysis is to embed the standard model (here the usual random effects method) within a class of alternative models which are at least plausible in the context of the data, and to compare the range of inferences given by this class. Of course the details will depend on which class is chosen — we do not claim that the class suggested in Section 2 contains all possible models or contains the "true" model in any absolute sense. Other classes could be taken, for example the Bayesian approach of Silliman (1997b) can be used in this way. Our idea is to keep as closely as possible to the usual maximum likelihood random effects model, but to add two further parameters which describe the study selection in a fairly transparent way. The specification of this in equation (2) below is a bit arbitrary, but some different versions of this equation have been tried and are found to make rather little difference to the range of inferences obtained.

2. Funnel Plots and Publication Bias

2.1. Model with Heterogeneity and Selection

Suppose that, of m studies to be reviewed, y_i is the estimated treatment effect (e.g. a log odds ratio) observed in the ith study. We assume that y_i is reported with standard error s_i. Allowing for between-study heterogeneity, we assume the variance components model

$$y_i = \mu_i + \sigma_i \epsilon_i, \quad \epsilon_i \sim N(0,1),\ \mu_i \sim N(\mu, \tau^2), \quad i = 1, \cdots, m. \tag{1}$$

Here, μ is the overall mean effect, τ^2 is the heterogeneity variance, σ_i^2 is the within study sampling variance, and ϵ_i and μ_i are assumed to be independent. We also assume that y_i and s_i are independent — this is exactly true if y_i is the sample mean of Normally distributed observations, but only approximately true in other cases such as when y_i and s_i are estimated from a 2×2 table.

A familiar approach to meta-analysis (*e.g.* Whitehead and Whitehead, 1991) is to assume that model (1) describes the studies in the review, and to estimate the parameters in the usual way. The problem of publication bias, however, arises when there are also other studies which have been carried out but which have not been published, or at least not included in the review. We model this by assuming that model (1) describes *all* the studies that have been done in the particular area of interest, but only *some* of them have been selected for review. Following Copas and Li (1997) and Copas (1999), we use a separate selection equation with a single correlation to model the selection (publication) process:

$$z_i = a + b/s_i + \delta_i \quad \delta_i \sim N(0,1), \quad corr(\epsilon_i, \delta_i) = \rho, \tag{2}$$

where the residuals (ϵ_i, δ_i) are assumed to be jointly Normal. The role of (2) is that

y_i is observed only when the latent variable $z_i > 0$.

The observed treatment effects are therefore modelled by the conditional distribution of y_i in (1), given that z_i in (2) is positive. For this reason we need to distinguish between $\sigma_i^2 = \mathrm{Var}(y_i|\mu_i)$ and s_i^2, which is an estimate of the conditional variance of y_i given that the study is published.

If $\rho = 0$ this is the model without publication bias; y_i and z_i are independent and so the outcome y_i has no effect on whether the paper is published or not. Notice that s_i appears on the right hand side of the z_i equation, so we are explicitly using the assumption that y_i and s_i are independent. But when $\rho > 0$, selected studies will have $z > 0$ and so δ, and consequently ϵ, are more likely to be positive, leading to a positive bias in y. Explicitly

$$E(y_i|z_i > 0, s_i) = \mu + \rho\sigma_i\lambda(a + b/s_i), \tag{3}$$

where $\lambda(\cdot)$ is Mill's ratio $\phi(\cdot)/\Phi(\cdot)$, ϕ and Φ being the density and distribution functions respectively of the standard Normal distribution. The shape of the points in Figure 1 suggests that this is what may be happening in the passive smoking example.

The parameters a and b in (2) control the marginal probability that a study with within-study standard deviation s is published. Parameter a controls the overall proportion published; parameter b controls how the chance of publication depends on study size. We expect b to be positive, so that very large studies (very small s) are almost bound to be accepted for publication, but only a proportion of the smaller ones will be accepted. And if $\rho > 0$ then those smaller studies that are accepted will tend to be those with large values of y. Since we do not observe how many unpublished studies there may have been, it is obvious that we cannot estimate a and b as unknown parameters in the usual way. We will show how a and b can be treated as free parameters in the sensitivity analysis.

Note that the models (1) and (2) are equivalent to the models

$$\begin{aligned}
&y_i = \mu + (\sigma_i^2 + \tau^2)^{1/2}\epsilon_i^*, \quad \epsilon_i^* \sim N(0,1),\\
&z_i = a + b/s_i + \delta_i \quad \delta_i \sim N(0,1), \quad corr(\epsilon_i^*, \delta_i) = \tilde{\rho}_i = \frac{\sigma_i}{(\tau^2 + \sigma_i^2)^{1/2}}\rho.
\end{aligned}$$

This notation makes clear that as there is only a single observation from each study the between and within study residuals can be combined into a single residual. It is now easy to write down the likelihood function as

$$\begin{aligned}
L(\mu, \rho, \tau, a, b) &= \sum_{i=1}^{m} [\log p(y_i|z_i > 0, s_i)]\\
&= \sum_{i=1}^{m} [\log p(y_i) + \log p(z_i > 0|y_i, s_i) - \log p(z_i > 0|s_i)]\\
&= \sum_{i=1}^{m} \left[-\frac{1}{2}\log(\tau^2 + \sigma_i^2) - \frac{(y_i - \mu)^2}{2(\tau^2 + \sigma_i^2)} - \log\Phi(u_i) + \log\Phi(v_i) \right],
\end{aligned} \tag{4}$$

where

$$u_i = a + \frac{b}{s_i}$$

and we have evaluated $p(z_i > 0|y_i, s_i)$ as $\Phi(v_i)$, with

$$v_i = \frac{u_i + \tilde{\rho}_i \frac{y_i - \mu}{(\tau^2 + \sigma_i^2)^{1/2}}}{(1 - \tilde{\rho}_i{}^2)^{1/2}}.$$

If we have sufficiently large sample sizes in each study, we can replace the nuisance parameters σ_i^2 by their sample estimates based on s_i^2. Since

$$\mathrm{Var}(y_i|s_i, z_i > 0) = \sigma_i^2(1 - c_i^2\rho^2)$$

where

$$c_i^2 = \lambda(u_i)(u_i + \lambda(u_i)),$$

we replace σ_i^2 in (4) by

$$\hat{\sigma}_i^2 = \frac{s_i^2}{1 - c_i^2\rho^2}.$$

Note that $\hat{\sigma}_i$ is not fixed, but depends on the other model parameters.

We can use (4) to find the profile likelihood for (a, b). We will find that this likelihood takes its maximum over a very flat plateau, confirming that there is not enough information to estimate the values of these parameters. However, for *given* (a, b), maximum likelihood estimates of (μ, ρ, τ) can be obtained by direct numerical maximization of (4).

The overall mean μ is the main parameter of interest in meta-analysis. For given (a, b), asymptotic inference about μ is summarized in the profile likelihood

$$L_{a,b}(\mu) = \max_{\rho,\tau|\mu,a,b} L(\mu, \rho, \tau, a, b).$$

Equating $2(L_{a,b}(\hat{\mu}) - L_{a,b}(\mu))$ to a percentile of the chi-squared distribution on one degree of freedom gives an approximate confidence interval for μ. If y_i is a log-odds ratio, the natural null hypothesis is that $\mu = 0$. The corresponding P-value in the asymptotic likelihood ratio test is

$$2\Phi(-[2(L_{a,b}(\hat{\mu}) - L_{a,b}(0))]^{\frac{1}{2}}).$$

Alternatively, the score test of this hypothesis is given by

$$t = \hat{\mu}\sqrt{(I_{11} - I_{12}I_{22}^{-1}I_{21})}, \tag{5}$$

where I is the Hessian matrix of the log likelihood function (4) with respect to (μ, ρ, τ) given (a, b), and I_{ij} is the related partition of I in terms of μ and (ρ, τ) respectively. The reciprocal of the quantity in the square root in (5) is the standard error of $\hat{\mu}$, which can be used to give a corresponding score-based confidence interval for μ. All these quantities are defined at $(\hat{\mu}, \hat{\rho}, \hat{\tau})$, and can be evaluated numerically from the likelihood (4). Table 1 lists the confidence intervals based on the likelihood ratio and score test for all three examples discussed in this paper — as expected, the two methods give quite similar results.

2.2. A Goodness of Fit Test for the Funnel Plot

If a specific (a, b) is to be entertained as a possible pair of parameters for the selection model, we need to check that the resulting model gives a reasonable fit to the funnel plot. We suggest testing this against the alternative that the some other pair (a^*, b^*) is better. If ρ or c_i^2 is small, it is easy to see from (3) that

$$E(y_i|z_i > 0, s_i, a^*, b^*) - E(y_i|z_i > 0, s_i, a, b) \approx c^* + \rho[\lambda(a^*) - \lambda(a)]s_i,$$

for some constant c^*. This suggests that local departures will be similar to adding a linear term in s_i to the expected value of y_i, and so testing a specific pair (a, b) is similar to testing $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ with restriction $\rho \geq 0$ in the model

$$y_i = \mu_i + \beta s_i + \sigma_i \epsilon_i, \tag{6}$$

$$z_i = a + b/s_i + \delta_i, \quad corr(\delta_i, \epsilon_i) = \rho. \tag{7}$$

This can be done directly by a likelihood ratio test similar to that discussed at the end of Section 2.1. For any given (a, b), write down the extended likelihood as expression (4) with the term βs_i added to μ, and evaluate the standardized likelihood ratio test for $\beta = 0$ in the usual way.

The naive model for meta-analysis without allowing for publication bias is equivalent to the original model (1) only, without the selection model (2), or to the full model with $\rho = 0$ or $a = \infty$. Thus the evidence for publication bias in the funnel plot can be tested as a special case of the above likelihood ratio test by setting $a = \infty$. This is equivalent to testing $\beta = 0$ in model (6). The likelihood ratio statistic is

$$\chi^2 = 2[\max_{\mu,\tau,\beta} \tilde{L}(\mu, \tau, \beta) - \max_{\mu,\tau} \tilde{L}(\mu, \tau, 0)] \tag{8}$$

where

$$\tilde{L}(\mu, \tau, \beta) = -\frac{1}{2} \sum_{i=1}^{m} [\log(\tau^2 + \sigma_i^2) + \frac{(y_i - \mu - \beta s_i)^2}{(\tau^2 + \sigma_i^2)}].$$

For the passive smoking example this gives $\chi^2 = 4.35$ on 1 *d.f.* or a P-value of 0.037, confirming the reasonably strong evidence for the presence of publication bias in this systematic review. Easier to calculate is the related score statistic

$$t = \hat{\beta}\sqrt{(\tilde{I}_{11} - \tilde{I}_{12}\tilde{I}_{22}^{-1}\tilde{I}_{21})}$$

where $\tilde{I}$ is the Hessian matrix of $\tilde{L}(\mu, \tau, \beta)$ and $\tilde{I}_{ij}$ is the related partition of $\tilde{I}$ in terms of β and (μ, τ). The formulae for $\tilde{I}$ are listed in Appendix 1. For the passive smoking data, $t = 2.13$ or a P-value of 0.033, almost the same as the likelihood ratio test.

Egger *et al.* (1997) proposed a simpler test for trend in a funnel plot, based on a simple regression of y_i on s_i. The above test with $a = \infty$ is equivalent to Egger's test in the special case of no heterogeneity, *i.e.* assuming that $\tau^2 = 0$ so that $\mu_i = \mu$ for all i.

The test we have proposed is motivated by local departures within the assumptions of the model, but may fail to detect mis-specification in the model itself. For a more general test, we could estimate the residuals

$$r_i(a,b) = \frac{y_i - E(y_i|z_i > 0, s_i, a, b)}{\sqrt(\tau^2 + s_i^2)}.$$

Under the model $r_i(a,b)$ is uncorrelated with s_i, and this could be tested non-parametrically. However, our examples suggest that this approach is not sufficiently powerful to be useful with the sample sizes (size of m) usually encountered in practice.

3. Sensitivity Analysis

Our idea is to develop a sensitivity analysis for inference about μ, allowing for a range of possible values of a and b. Let P_s be the marginal selection probability of a typical study with standard error s, so that

$$P_s = P(z > 0|s, a, b) = \Phi(a + b/s).$$

For example, if we set $(a,b) = (-1, .5)$, P_s takes values 37.46% and 99.99% for the smallest and largest observed studies in the passive smoking review. This means that almost all studies similar to the largest one in the review will be published, but only one third of studies similar to smallest one will be published. More generally, the relationship between P_s and s is illustrated in Figure 2. The range of values of s used in this graph is the range observed in the 37 studies, from 0.10 (most accurate study) to 0.73 (least accurate study). The top curve has $a = 0$ and $b = 2.0$, showing that almost all studies will be published (P_s close to 1). The bottom curve has $a = -3$ and $b = 0.1$. Here, the published studies are highly selective, with a low publication probability even for small values of s. These curves seem rather extreme, more plausible curves probably lie somewhere in between, *e.g.* the curve for $a = -1$ and $b = .5$ discussed above. This suggests that our analysis should look at values of a in the range -3 to 0, and values of b in the range 0.1 to 2.0.

For any pair of given values of (a,b), μ can be estimated by maximum likelihood as in Section 2. A contour diagram of $\hat{\mu}$ over these ranges of a and b is presented in Figure 3(i). At the top right (corresponding to the top curve in figure 2, meaning very little selection bias) the estimate $\hat{\mu}$ is 0.22 (relative risk is 1.24), which is the same estimate as the one estimated from the model without selectivity. But as we move away from this corner of the graph the log odds ratio falls sharply, down to only 0.04 (*i.e.* only a 4% excess risk) at the lower left (corresponding to the lower curve in figure 2, high potential for publication bias).

The values of $\hat{\mu}$ in Figure 3(i) have to be judged in the light of the goodness of fit to the funnel plot. Minus one times the profile likelihood for (a,b) is shown in Figure 4 (the perspective diagram is easier to see this way up). Maximum likelihood is achieved towards the "front" of the plot, but in this region the likelihood is very flat, suggesting a wide range of values of (a,b) which give an equally good fit. But towards the "back" of the plot the likelihood falls quite sharply, suggesting that values of (a,b) which give publication probabilities close to one are not acceptable. These are just the values of (a,b) which give $\hat{\mu}$ close to the the crude weighted average estimate. There is a narrow diagonal band across the middle of the (a,b) range where $\hat{\rho}$ is close to 1, leading to numerical difficulties in the evaluation of this profile likelihood — this explains the irregularities apparent in Figure 4.

An alternative approach is to use the likelihood ratio test proposed in the last section, and calculate the P-value as a function of (a, b). Contours of these P-values are shown in Figure 3 (ii). Using 5% as a conventional threshold, this identifies an area to the upper right corner within which the model gives a significantly poor fit. Again these are the values of (a, b) giving high publication probabilities. Comparing Figures 3(i) and 3(ii) we see that these systems of contours are very roughly parallel, suggesting that $\hat{\mu}$ can be plotted directly against the P-value. This is in Figure 3(iii). The points in this plot correspond to a rectangular grid of values of (a, b) over the chosen range. Of course they do not exactly lie on a single curve as the contours are not exactly parallel, but this plot does indicate how $\hat{\mu}$ decreases as the quality of fit improves. The estimated log odds ratio is always less than .20 when the P-value is greater then 0.05, *i.e.* for an acceptable fit to the funnel plot the estimate of excess risk is at most 22%. Now if a null hypothesis is true the P-values are uniformly distributed between 0 and 1, and so have expectation 0.5. For this "average P-value", the risk excess is only 14%.

The fit to the funnel plot given by different values of (a, b) is also illustrated by the dashed lines in Figure 5. These curves are the fitted values calculated from (3). Two values of (a, b) give a good, and almost indistinguishable, fit. The third is clearly unacceptable, consistent with the above profile likelihood and P-value plots.

In order to interpret this information further, and to aid its description in tabular form, it is useful to try to calibrate (a, b) into a single and more interpretable parameter. The fact that the contours in Figure 3(i) are roughly parallel lines means that $\hat{\mu}$ depends only on a single linear combination of a and b. Calculating the slope of the contours near the centre of the plot gives the best such combination to be approximately

$$a + 5b = a + b/0.2.$$

But $P_s = \Phi(a + b/s)$, and so $\hat{\mu}$ depends on $P_{0.2}$, the marginal publication probability for a study with $s = 0.2$ (an accuracy within the observed range). We call $P_{0.2}$ the *canonical probability*. The relationship between $P_{0.2}$ and $\hat{\mu}$ is shown in Figure 3(iv). The crude estimate of 24% excess risk is given only when $P_{0.2} = 1$, but if $P_{0.2} = 0.7$ (70% of studies with $s = 0.2$ published) the risk excess is roughly halved. The value of s defining the canonical probability can be estimated directly from the Hessian of the likelihood, as pointed out in Appendix 2.

Table 1 summarizes the sensitivity analysis for the passive smoking review. For each value of the canonical probability the table shows $\hat{\mu}$ along with its approximate 95% confidence limits. The limits are calculated by both methods discussed in Section 2.1. Also shown are the P-values for the fit to the funnel plot. Since, as explained, these quantities are not exactly functions of $P_{0.2}$, some judgement has been used in choosing the inference for a single value of (a, b) along the appropriate contour. The last line in the first panel of Table 1 shows that the standard random effects estimate of 0.22 is not acceptable on account of the small P-value in the penultimate column. As commented earlier, $\hat{\mu}$ needs to be at most 0.20 for the P-value to rise above 5%.

The smaller the value of P_{s_i} the larger the number of unpublished studies being imputed by the model. A rough estimate of the total number of unpublished studies is

$$\sum_{i=1}^{m} \frac{1 - P_{s_i}}{P_{s_i}}.$$

The last column of table 1 gives these estimates for the values of (a, b) we have selected. These numbers should not be interpreted too literally, but may help in a subjective assessment of the different values of a and b.

This way of looking at systematic reviews is a sensitivity analysis and does not claim to produce a definitive estimate of μ. However, the conclusion that the crude weighted average relative risk in the passive smoking review is an overestimate, possibly a substantial overestimate, seems inescapable. If the use of the "expected P-value" is accepted, then the analysis suggests that the most plausible value of the excess risk may be around 14%, down to almost half the published estimate. The evidence for whether there is a risk increase at all then becomes questionable, as the confidence interval then just includes zero. This is a very different result from the standard analysis, and yet the predicted 28 unpublished studies (last column of Table 1) is surely not excessive, implying that this systematic review has captured about $35/(35+28) = 56\%$ of all studies that have been attempted in this area.

4. More Numerical Examples

He *et al* (1999) reported another systematic review on the risks of passive smoking, but looking at coronary heart disease rather than lung cancer. In this review there were 10 cohort and 8 case-control epidemiological studies. Figure 6 is the funnel plot of the log odds ratio against the standard error. Again there is a clear sign of publication bias, even if we disregard the study with the largest standard error as an outlier (this study is much smaller than the others and in fact is given very little weight in the analysis). The likelihood ratio test for the presence of publication bias gives a P-value of 0.001 (or 0.002 if the smallest study is excluded).

The conventional estimate of μ (ignoring selection bias) is 0.25, or an estimated excess risk of 28%. The sensitivity analysis again suggests that this is an overestimate. With 5% as a conventional threshold for the fit to the funnel plot, the estimate of μ is at most 0.20 (excess risk 22%), see Figure 7. The fall in the value of $\hat{\mu}$ as the P-value increases from zero is more marked in this example, reflecting the greater evidence for publication bias shown in the funnel plot. Table 1 summarizes the sensitivity analysis. It shows that the estimate of μ is only about 0.17 (excess risk 18%) when the canonical probability $P_{.125}$ is 0.7.

Our third example is a systematic review considering the effect of selective decontamination of the digestive tract on the risk of respiratory tract infection; patients in intensive care units were randomized to receive treatment by a combination of non-absorbable antibiotics or to receive no treatment (Smith, *et al*, 1995). Here 22 trails are reviewed. The funnel plot in Figure 8 shows a very strong trend for publication bias. The likelihood ratio statistic gives $\chi^2 = 20.9$. The results of our sensitivity analysis are summarized in Table 1 and Figure 9. Here the fall in $\hat{\mu}$ as the P-value increases is even more marked. The crude estimate of μ is 1.28, but it is at most 0.8 if we require an acceptable fit to the funnel plot in the sense that the P-value be not be less than 0.05.

5. Discussion

In our examples we have illustrated the data by plotting y against s, but other graphs are possible. Funnel plots are often shown with the axes the other way round, or with s replaced by study sample size shown on a log scale. The Galbraith plot (plot of y_i/s_i against $1/s_i$ — see Galbraith, 1988) is a good alternative — the standard fixed effect model then gives a constant scatter and the estimate of μ is just least squares through the origin. We have chosen to plot y against s since we want to superimpose the fitted values given by equation (3) which are more naturally thought of as a function of s.

We have used the examples in the paper as illustrations of our method, and have ignored other aspects of data quality which in practice can and should be examined. It is not uncommon to find inconsistencies when the same study is used by different reviewers. Lee (1999) discusses some simple data checks, and points out inconsistencies in at least one of the passive smoking studies. In systematic reviews of epidemiological studies there can be major problems in ensuring comparability in measures of exposure and outcome. Hackshaw *et al.* (1997) give a careful discussion of several possible confounding factors.

Publication bias is a special case of the problem of non-ignorable missing data. In the jargon of the missing data literature (*e.g.* Little and Rubin, 1987), observations are "missing at random" if the event that an observation is missing is conditionally independent of the actual value of that observation, given all information that has been observed on that individual or experimental unit. If the research studies being reviewed are thought of as the observations, then in our model the unpublished studies are missing at random only if $\rho = 0$. There is a very large literature on missing data problems, and some of the ideas and discussion are very relevant to meta-analysis.

We commented in Section 1 that correcting for publication bias is only possible if we are prepared to make unverifiable assumptions. Even sensitivity analyses are based on assumptions — here we have made the crucially important assumption implicit in (1), that there is no intrinsic connection between y_i and s_i. Only with this assumption can we conclude that a trend in the funnel plot is indicative of publication bias. This assumption cannot be checked from the data themselves, but needs to be assessed in the context of the subject matter. Conceivably, the studies with small s, being the larger trials, may be better organised and controlled and so may tend to give better results (larger y). Conversely, perhaps studies with large s, being the smaller trials, may be run by a smaller and more committed staff and so may give the better results for that reason. This assumption is essentially the null hypothesis that $\beta = 0$ in model (6), and so corresponds to the test for the presence of publication bias which we have proposed in Section 2. If there are other reasons why $\beta \neq 0$ then it is seems impossible to say anything very useful at all about the effects of publication bias.

The estimates in Table 1 of the number of unpublished studies are in a sense the minimum number of unselected studies needed to explain the degree of publication bias being entertained. The crucial question is not how many unpublished studies are in the pool from which the ones for review are sampled, but how this sampling is done. If the selection is done in two stages, first model (1) selects a set of candidate studies, and then the actual studies in the review are sampled *randomly* from this set, then the sampling fraction used at the second stage is of course irrelevant.

We have assumed standard asymptotic properties of maximum likelihood throughout. The Normality of y and its independence from s is at best an approximation, and the fact that the maximum likelihood estimate of ρ can be attained at a boundary point of the parameter space raises doubts about the likelihood ratio and score tests proposed in Section

2. However, because of the essential indeterminacy in (a, b), we would argue that numerical accuracy at any given (a, b) is not important. All we have attempted to do is to suggest that meta-analysis needs a model which explicitly allows for publication bias, and to see if a substantially lower estimate of μ is needed for such a model to give a reasonable fit to the funnel plot. Arguably, the question of whether this selection model is "correct" is not the issue — the fact that a reasonably plausible model exists is enough to cast doubt on the conventional analysis.

References

Copas,J.B. (1999). What works?: selectivity models and meta-analysis. *J. R. Statist. Soc. A*, **162**, 95-109.

Copas,J.B. and Li,H.G. (1997). Inference for non-random sample (with discussion). *J. R. Statist. Soc. B*, **59**, 55-95.

Copas,J.B. and Shi, J.Q. (2000). Reanalysis of epidemiological evidence on lung cancer and passive smoking. *British Medical Journal*, **320**, 417-418.

Dear,H.B.G. and Begg,C.B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science*, **7**, 237-245.

Egger,M., Smith,G.D., Schneider,M. and Minder,C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, **315**, 629-634.

Galbraith,R. (1988) A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine*, **8**, 889-894

Givens,G.H., Smith,D.D. and Tweedie,R.L. (1997) Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate (with discussion). *Statistical Science*, **12**, 244-245.

Hackshaw,A.K., Law M.R. and Wald,N.J. (1997) The accumulated evidence on lung cancer and environmental tobacco smoke. *British Medical Journal*, **315**, 980-988.

He,J. *et al*, (1999). Passive smoking and the risk of coronary heart disease – A meta analysis of epidemiologic studies. *The New England J. of Medicine*, **340**, 920-926.

Hedges,L.V. (1984). Estimation of effect size under nonrandom sampling: the effects of censoring studies yielding statistically insignificant mean differences. *J. of Educational Statistics*, **9**, 61-85.

Hedges,L.V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, **7**, 227-236.

Iyengar,S. and Greenhouse,J.B. (1988). Selection models and the file drawer problems (with discussion). *Statistical Science*, **3**, 109-135.

Lee, P. (1999) Simple methods for checking for possible errors in reported odds ratios, relative risks and confidence intervals. *Statistics in Medicine*, **18**, 1973-1981.

Little,R.J.A. and Rubin,D.B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Poswillo, D. *et al* (1998). *Report of the Scientific Committee on Tobacco and Health.* Department of Health, *et al.* London: The Stationery Office.

Silliman,N.P. (1997a). Nonparametric classes of weight functions to model publication bias. *Biometrika*, **84**, 909-918.

Silliman,N.P. (1997b). Hierarchical selection models with applications in meta-analysis. *J. of American Statistical Association*, **92**, 926-936.

Smith,T.C., Spiegelhalter,D.J. and Thomas,A. (1995). Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine*, **14**, 2685-2699.

Sutton,A.J., Duval,S.J., Tweedie,R.L., Abrams,K.R. and Jones,D.R. (1999) The impact of publication bias on meta-analyses within the cochrane database of systematic review. *Technical Report, Dept of Epidemiology and Public Health, U. of Leicester, UK.*

Taylor,S.J. and Tweedie,R.L. (1998a) A non-parametric "trim and fill" method of assessing publication bias in meta-analysis. *Technical Report, Dept of Statistics, Colorado State University, USA.*

Taylor,S.J. and Tweedie,R.L. (1998b) Trim and fill: A simple funnel plot based methods of testing and adjusting for publication bias in meta-analysis. *Technical Report, Dept of Statistics, Colorado State University, USA.*

Whitehead, A. and Whitehead, J. (1991) A general parametric approach to the meta analysis of randomized clinical trials. *Statistics in Medicine*, **10**, 1665-1677.

Acknowledgement

One of us (JQS) is supported by a grant from the *ESRC*.

Appendix

A1: Formula for the Hessian matrix $\tilde{I}$ in Section 2.2

The 3×3 matrix $\tilde{I}$ is

$$\tilde{I} = \begin{bmatrix} L_{\beta\beta} & L_{\beta\mu} & L_{\beta\tau} \\ L_{\mu\beta} & L_{\mu\mu} & L_{\mu\tau} \\ L_{\tau\beta} & L_{\tau\mu} & L_{\tau\tau} \end{bmatrix}$$

where

$$L_{\beta\beta} = \sum_{i=1}^{m} \frac{s_i^2}{\tau^2 + s_i^2}, \quad L_{\beta\mu} = \sum_{i=1}^{m} \frac{s_i}{\tau^2 + s_i^2}, \quad L_{\beta\tau} = 2\sum_{i=1}^{m} \frac{(y_i - \mu - \beta s_i)\tau s_i}{(\tau^2 + s_i^2)^2},$$

$$L_{\mu\mu} = \sum_{i=1}^{m} \frac{1}{\tau^2 + s_i^2}, \quad L_{\mu\tau} = 2\sum_{i=1}^{m} \frac{(y_i - \mu - \beta s_i)\tau}{(\tau^2 + s_i^2)^2}, \quad L_{\tau\tau} = \sum_{i=1}^{m} \frac{\tau^2 - s_i^2}{(\tau^2 + s_i^2)^2}.$$

A2. Identifying the Canonical Probability P_s

Let $\boldsymbol{\gamma}^T = (a, b)$ and $\boldsymbol{\theta}^T = (\mu, \tau, \rho)$, and let $L(\boldsymbol{\gamma}, \boldsymbol{\theta})$ be the log likelihood in (4). Choose a representative value of $\boldsymbol{\gamma}$ near the centre of the chosen range of (a, b), and denote this by

$\boldsymbol{\gamma}_0^T = (a_0, b_0)$. Let $\boldsymbol{\theta}_0$ be the corresponding MLE given by $\boldsymbol{L}_\theta(\boldsymbol{\gamma}_0, \boldsymbol{\theta}_0) = 0$, where $\boldsymbol{L}_\theta$ is the first derivative of L with respect to $\boldsymbol{\theta}$. Expanding the log likelihood we have

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\theta}) &= L(\boldsymbol{\gamma}_0, \boldsymbol{\theta}_0) + \boldsymbol{L}_\gamma(\boldsymbol{\gamma}_0, \boldsymbol{\theta}_0) - \frac{1}{2}\Big[(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T \boldsymbol{H}_{11}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) \\ &\quad +2(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{H}_{21}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{H}_{22}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\Big] \end{aligned}$$

where $\boldsymbol{L}_\gamma$ is the first derivative of L with respect to $\boldsymbol{\gamma}$, $\boldsymbol{H}$ is Hessian Matrix with respect to $(\boldsymbol{\gamma}, \boldsymbol{\theta})$, and $\boldsymbol{H}_{ij}$ is the related partition of $\boldsymbol{H}$ in terms of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ respectively. The MLE of $\boldsymbol{\theta}$ for an arbitrary $\boldsymbol{\gamma}$ close to $\boldsymbol{\gamma}_0$ is therefore given by

$$\hat{\boldsymbol{\theta}} \simeq \boldsymbol{\theta}_0 - \boldsymbol{H}_{22}^{-1} \boldsymbol{H}_{21}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T.$$

Let (g_0, g_1) be the first row of $\boldsymbol{H}_{22}^{-1} \boldsymbol{H}_{21}$. Then,

$$\begin{aligned} \hat{\mu} &\simeq \mu_0 - [g_0(a - a_0) + g_1(b - b_0)] \\ &= \mu_0 + g_0 a_0 + g_1 b_0 - g_0(a + \frac{g_1}{g_0} b) \end{aligned}$$

and so $\hat{\mu}$ depends on P_s where $s = g_0/g_1$.

Table 1. The estimates of μ, 95% confidence intervals of μ from likelihood ratio (CI1) and score functions (CI2), the P-value for fit to the funnel plot, and estimated numbers (#) of unpublished studies for different canonical probabilities (CP) .

	CP	$\hat{\mu}$	CI1	CI2	P-value	#
Passive smoking and lung cancer data						
	0.6	.11	(-.02, .23)	(.00, .21)	.594	42
	0.7	.12	(.00, .25)	(.02, .22)	.508	28
$P_{.2}$	0.8	.14	(.02, .26)	(.04, .23)	.355	15
	0.9	.16	(.05, .28)	(.06, .26)	.203	8
	1.0	.22	(.12, .31)	(.12, .32)	.037	
Passive smoking and coronary heart disease data						
	0.6	.17	(.11, .25)	(.11, .24)	.515	51
	0.7	.17	(.11, .26)	(.12, .24)	.404	27
$P_{.125}$	0.8	.18	(.11, .26)	(.12, .24)	.316	18
	0.9	.20	(.13, .25)	(.13, .26)	.471	9
	1.0	.25	(.16, .41)	(.07, .43)	.001	
Respiratory tract infaction data						
	0.6	.67	(.40, .99)	(.40, .95)	.098	54
	0.7	.71	(.43, 1.04)	(.41, 1.00)	.058	35
$P_{.4}$	0.8	.77	(.47, 1.12)	(.45, 1.10)	.022	18
	0.9	.95	(.61, 1.27)	(.57, 1.23)	.004	8
	1.0	1.28	(.92, 1.73)	(.89, 1.68)	$<$.001	

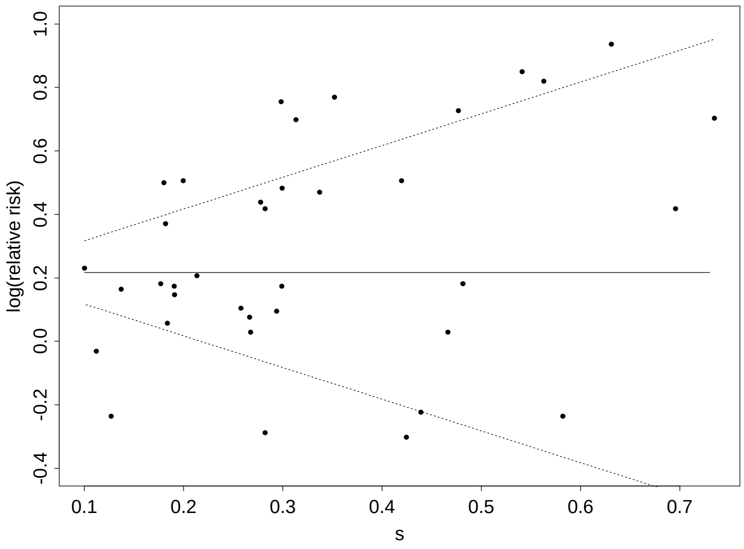

Figure 1. Passive smoking and lung cancer data: Funnel plot; the solid line represents $\hat{\mu} = .22$, the estimate without selectivity; the dotted lines represent $\mu \pm s$.

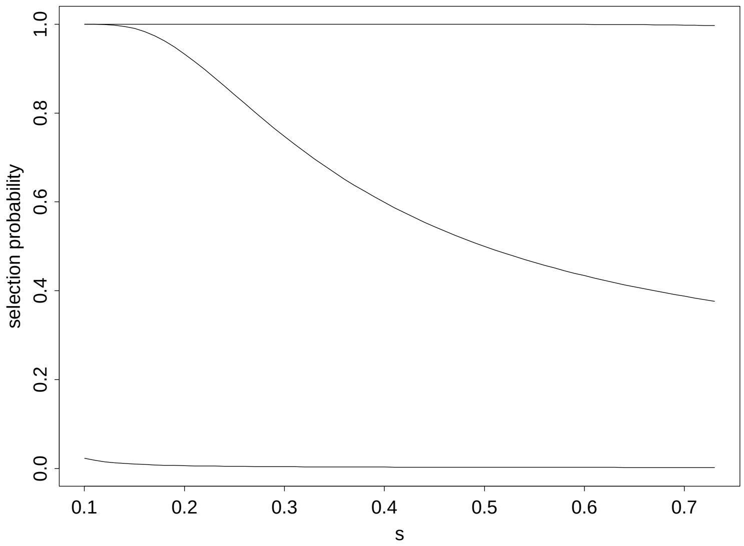

Figure 2. Passive smoking and lung cancer data: The selection probability for $(a, b) = (0, 2)$ (the top curve), $(a, b) = (-1, .5)$ (the middle curve) and $(a, b) = (-3, .1)$ (the bottom curve).

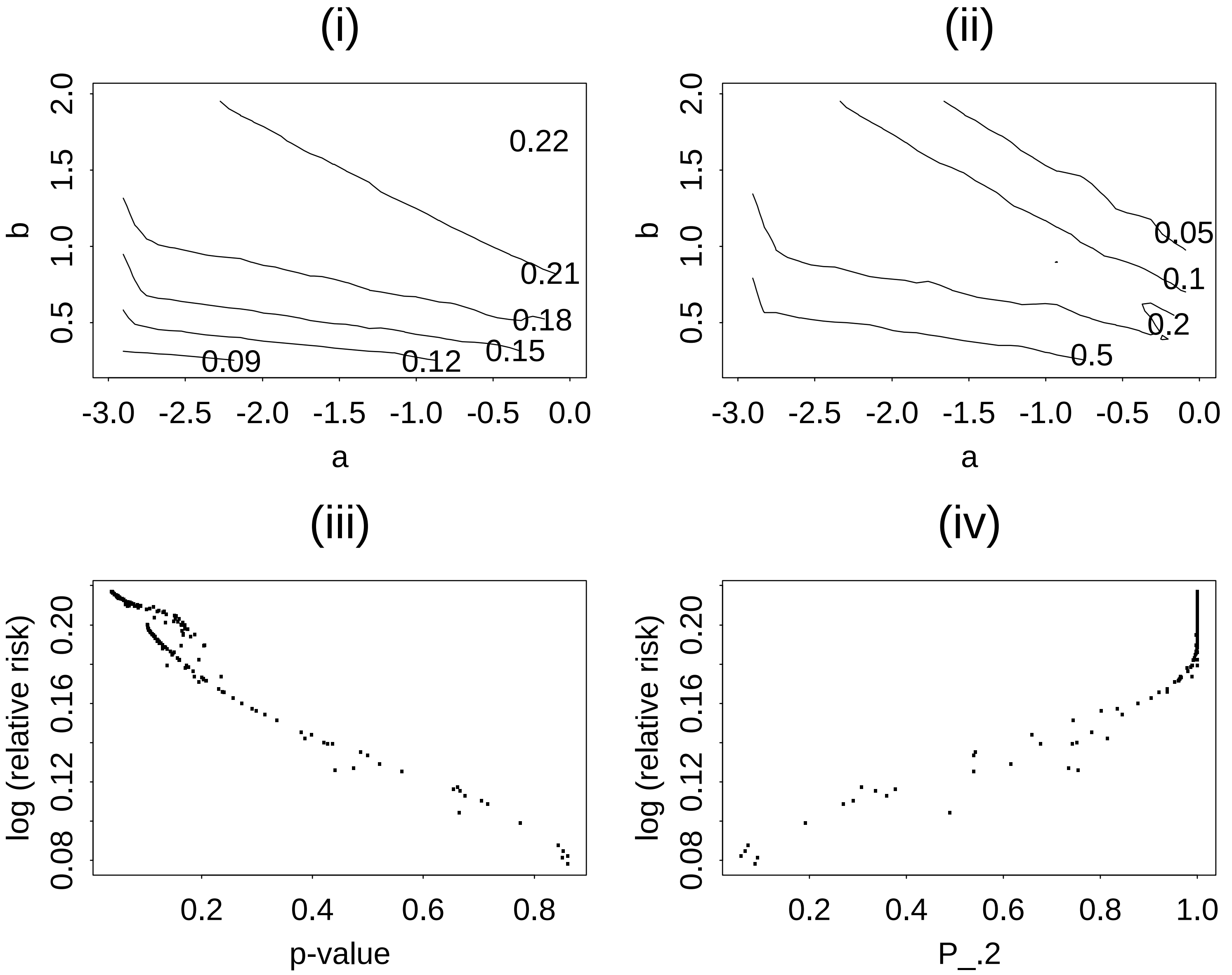

Figure 3. Passive smoking and lung cancer data: (i) Contours of $\hat{\mu}$; (ii) Contours of P-values of $H_0:\ \beta = 0$ from likelihood ratio test; (iii) $\hat{\mu}$ against the P-value; (iv) $\hat{\mu}$ against canonical probability $P_{.2}$.

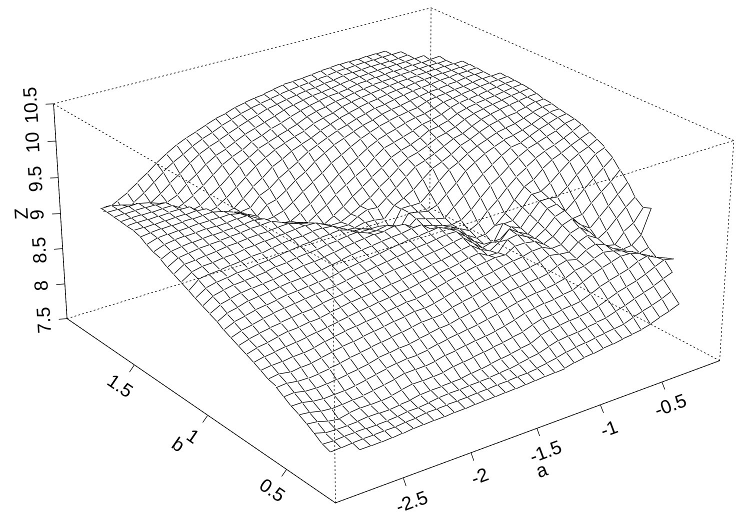

Figure 4. Passive smoking and lung cancer data: Perspective diagram of negative profile log-likelihood for (a, b).

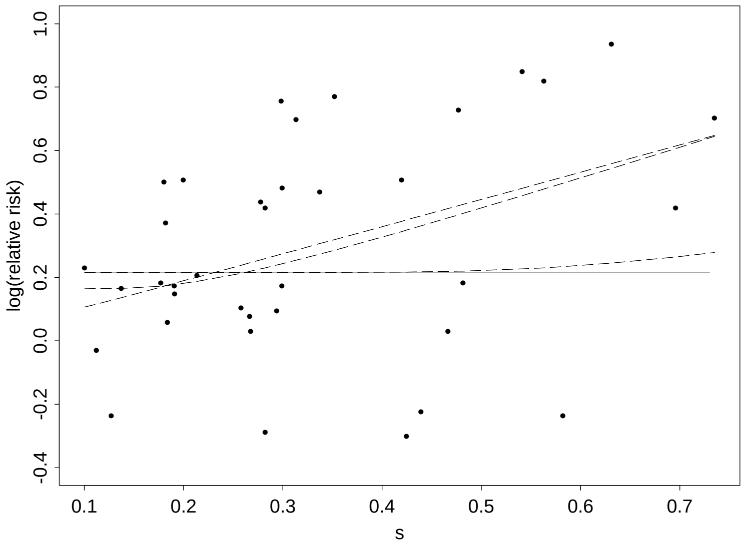

Figure 5. Passive smoking and lung cancer data: Funnel plot; the solid line represents $\hat{\mu} = .22$, the estimate without selectivity; the dashed lines represent fitted value for given (a, b), which $(a, b, \hat{\mu})$ are equal to $(-.5, 1.5, .22)$, $(-1, .5, .17)$ and $(-2.5, .2, .07)$ respectively.

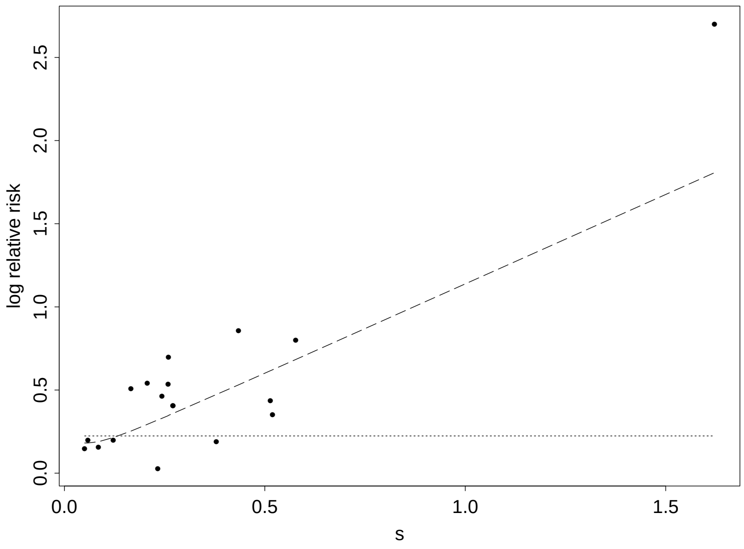

Figure 6. Passive smoking and coronary heart disease data: Funnel plot; the dotted line represents the estimate without selectivity $\hat{\mu} = .25$; the dashed lines represent fitted values given $(a, b) = (-.76, .2)$, the related estimate is $\hat{\mu} = .18$.

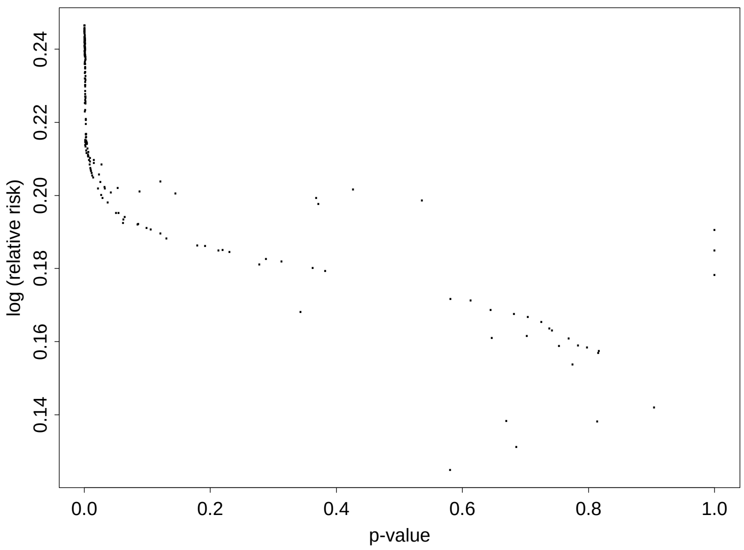

Figure 7. Passive smoking and coronary heart disease data: $\hat{\mu}$ against the P-value.

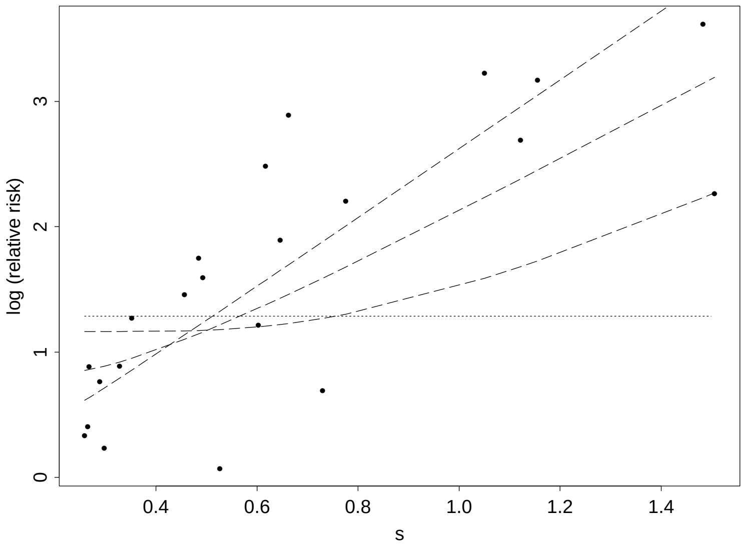

Figure 8. Respiratory tract infection data: Funnel plot; the dotted line represents the estimate without selectivity $\hat{\mu} = 1.28$; the dashed lines represent fitted values for given (a, b), which $(a, b, \hat{\mu})$ are equal to $(-.5, 1.5, 1.16)$, $(-.5, .5, .81)$ and $(-2.5, .5, .34)$ respectively.

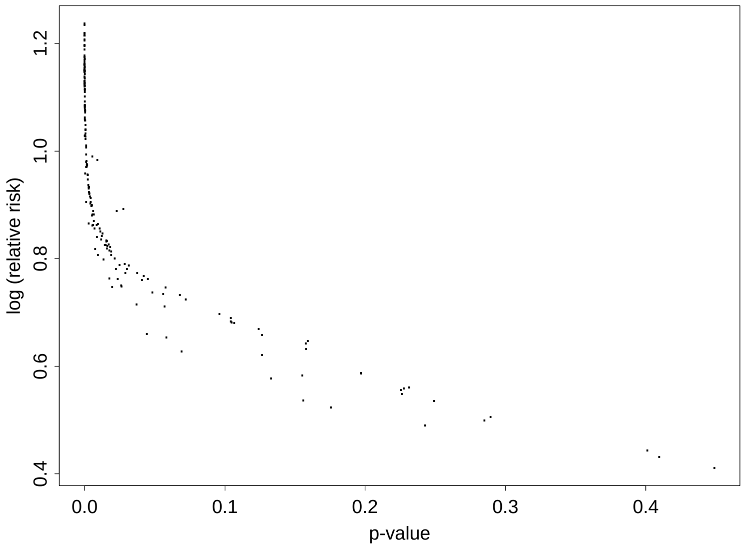

Figure 9. Respiratory tract infection data: $\hat{\mu}$ against the P-value.