11

Model Checking in Meta-Analysis

Wolfgang Viechtbauer

CONTENTS

11.1 Introduction

In previous chapters, methods have been described to compute and model various outcome or effect size measures, such as risk differences, (log transformed) risk/odds ratios, raw or standardized mean differences, and correlation coefficients. The observed values of such measures may reflect the size of a treatment effect, the degree to which a risk factor is related to the chances of being afflicted by (or the severity of) a particular disease, or more generally the size of group differences. Some measures, such as the correlation coefficient, simply reflect the degree to which two variables of interest are (linearly) related to each other.

The process of analyzing such data involves fitting one or more models to the observed outcomes,* based on which we can draw conclusions about the effectiveness of a treatment, the relevance of a risk factor, the degree to which groups differ, the strength of the association between two variables, and so on. In addition, it is typically of interest to examine whether the phenomenon being studied (e.g., the treatment effect) is relatively homogeneous across studies or varies, possibly as a function of one or more variables that can account for this heterogeneity. However, the models used in such analyses make various assumptions. In practice, assumptions may be violated, which in turn may affect the statistical properties of the inferential methods used to draw conclusions from the data at hand. It is therefore important to carefully consider to what extent the various assumptions may be violated and what impact this may have on the results and conclusions.

Similarly, when we fit a model to our data, we make the implicit assumption that the model represents an adequate approximation to some underlying data generating process. Naturally, reality is more complex than any model we can envision, but gross mismatch between the data and model should warn us that we are far from providing an adequate description of how the data may have arisen. Therefore, assessing model fit should be an essential step in any analysis. For some aspects of the models, this can be done by means of a statistical test. An examination of the residuals and standardized versions thereof can also provide clues about the presence of an incongruity between the data and the assumed model and/or may indicate that the model is not appropriate for certain data points.

Finally, it is important to examine whether one or more studies exert a disproportionally large influence on the conclusions of a meta-analysis. If we find that certain findings hinge on only one or two studies being present in our dataset, then this may call into question the robustness of the findings, in which case the corresponding conclusions should be framed more cautiously. An important tool in this context is to examine how the removal of studies from the dataset would alter the results. We will therefore consider a variety of diagnostic measures on the basis of this idea.

11.2 Models Assuming Normal Sampling Distributions

In this section, some of the meta-analytic models introduced in previous chapters will be reviewed, with emphasis on the common-, random-, and (mixed-effects) meta-regression models. The models considered here are all based on the assumption that the observed outcomes represent draws from normal sampling distributions (models that assume other types of sampling distributions will be briefly discussed in Section 11.7). Moreover, the models are not tied to any particular outcome or effect size measure. Therefore, in general, let y_i denote the observed value of the chosen outcome measure (e.g., log odds ratio) in the ith study and let θ_i denote the value of the corresponding (unknown) true outcome (e.g., the true log odds ratio).

* As described in Chapter 5, it is also possible to model the raw data directly where available (or where it can be reconstructed based on the available information) using appropriate generalized linear mixed-effects models. We will return to a discussion of such models at the end of this chapter.

11.2.1 Common-Effect Model

The first model we will consider is the common-effect model, which is given by

$$y_i = \theta + e_i, \tag{11.1}$$

where θ denotes the underlying true value of the outcome measure and $e_i \sim N(0, s_i^2)$. Therefore, as the name implies, the model assumes that the true outcomes are equal (homogeneous) across studies (i.e., $\theta_i = \theta$ for $i = 1, \ldots, k$). Moreover, the sampling error in the ith study (i.e., e_i) is assumed to be normally distributed with sampling variance equal to s_i^2, which in turn we consider to be a known and fixed quantity. Finally, assuming independence between studies implies $\text{Cov}[e_i, e_{i'}] = 0$ for $i \neq i'$.

11.2.2 Random-Effects Model

The common-effect model can be considered a special case of the more general random-effects model, where potential heterogeneity in the true outcomes is accounted for by adding a random effect for each study to the model. In particular, the random-effects model is given by

$$y_i = \mu + u_i + e_i, \tag{11.2}$$

where μ denotes the average true outcome and $u_i \sim N(0, \tau^2)$ is a normally distributed random effect by which the true outcome in the ith study differs from the average true outcome (so that $\theta_i = \mu + u_i$). Therefore, τ^2 denotes the amount of variance (or "heterogeneity") in the true outcomes (hence, if $\tau^2 = 0$, then the random-effects model simplifies to the common-effect model, so that $\mu \equiv \theta$). Assumptions about e_i are as described previously, with the addition that we assume independence between different u_i values (and hence, $\text{Cov}[u_i, u_{i'}] = 0$ for $i \neq i'$) and between u_i and e_i (which implies $\text{Cov}[u_i, e_i] = 0$).

11.2.3 Meta-Regression Model

An alternative approach to account for heterogeneity in the true outcomes is to explicitly model such differences by means of one or more predictor (or "moderator") variables. This leads to the meta-regression model, which is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + e_i, \tag{11.3}$$

where β_1 through β_q are model coefficients that denote how the true outcome changes for a one-unit increase in the corresponding moderator variables x_{i1} through x_{iq} and β_0 denotes the model intercept, which corresponds to the true outcome when all moderator variables take on the value 0.

11.2.4 Mixed-Effects Model

Finally, analogous to the random-effects model, the meta-regression model can also be extended by the addition of a random effect for each study. Doing so yields the mixed-effects model (also called random-effects meta-regression model), which is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + u_i + e_i, \tag{11.4}$$

with $u_i \sim N(0,\tau^2)$ as before, except that τ^2 should now be interpreted as the amount of variance in the true outcomes that is not accounted for by the moderator variables included in the model (hence, τ^2 is often described as the amount of "residual heterogeneity").

11.2.5 Model Fitting

Methods for fitting the various models described above were discussed in earlier chapters but will be briefly restated here for completeness sake and to fix notation. We start with the mixed-effects model and then treat the remaining models as special cases thereof. Matrix notation is used, as it provides a compact way of writing out the equations. When it is helpful for understanding, algebraic expressions are also provided.

The mixed-effects model implies that $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V} + \tau^2\boldsymbol{I})$, where $\boldsymbol{y}$ is a $k \times 1$ column vector with the observed outcomes y_1 through y_k, $\boldsymbol{X}$ is a $k \times (q+1)$ matrix containing the values of the moderator variables (with the constant 1 in the first column for the model intercept), $\boldsymbol{\beta}$ is a $(q+1) \times 1$ column vector with the model coefficients $\beta_0, \beta_1, \ldots, \beta_q$, $\boldsymbol{V}$ is a diagonal matrix with the s_i^2 values along the diagonal, and $\boldsymbol{I}$ is a $k \times k$ identity matrix.

To fit the mixed-effects model, we must first estimate τ^2 using one of the various estimators that have been described in the literature for this purpose (e.g., Thompson and Sharp, 1999; Viechtbauer et al., 2015) and that are reviewed in Chapter 4. For the purposes of this chapter, we will just consider a relatively simple method-of-moments estimator, which is given by

$$\hat{\tau}^2 = \frac{Q_E - (k-q-1)}{\text{trace}[\boldsymbol{P}]}, \tag{11.5}$$

where $Q_E = \boldsymbol{y'Py}$ and $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X'V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X'V}^{-1}$. In case $\hat{\tau}^2$ is negative, the estimate is set to 0.

Once $\hat{\tau}^2$ has been calculated, estimates of the model coefficients in $\boldsymbol{\beta}$ can be obtained using the weighted least squares estimator

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'WX})^{-1}\boldsymbol{X'Wy}, \tag{11.6}$$

where $\boldsymbol{W} = (\boldsymbol{V} + \hat{\tau}^2\boldsymbol{I})^{-1}$. The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ can be estimated with

$$\text{Var}[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X'WX})^{-1}. \tag{11.7}$$

Taking the square root of the diagonal elements of $\text{Var}[\hat{\boldsymbol{\beta}}]$ then yields standard errors of the model coefficients (i.e., $SE[\hat{\beta}_0], \ldots, SE[\hat{\beta}_q]$). Note that (11.7) ignores the uncertainty in the estimate of τ^2, which is discussed in more detail in Chapter 4.

By comparing the test statistic $z_j = \hat{\beta}_j / SE[\hat{\beta}_j]$ against appropriate percentiles of a standard normal distribution (e.g., ±1.96 for a two-sided test at $\alpha = 0.05$), we can test H_0: $\beta_j = 0$, that is, whether there is a significant relationship between a moderator variable and the outcomes. Analogously, an approximate 95% confidence interval (CI) for β_j can be constructed with $\hat{\beta}_j \pm 1.96 SE[\hat{\beta}_j]$. Simultaneous tests of multiple coefficients can also be conducted by computing

$$Q_M = \hat{\boldsymbol{\beta}}'_{[2]}(\text{Var}[\hat{\boldsymbol{\beta}}]_{[2]})^{-1}\hat{\boldsymbol{\beta}}_{[2]}, \tag{11.8}$$

where $\hat{\boldsymbol{\beta}}_{[2]}$ is a column vector containing the m coefficients to be tested and $\text{Var}[\hat{\boldsymbol{\beta}}]_{[2]}$ denotes an $m \times m$ matrix with the corresponding rows and columns from (11.7). Under the null hypothesis that the true values of the coefficients tested are all equal to zero, Q_M follows (approximately) a chi-square distribution with m degrees of freedom. A common application of (11.8) is to test all coefficients excluding the intercept (i.e., H_0: $\beta_1 = \cdots = \beta_q = 0$), which (analogous to the omnibus F-test in multiple regression) can be used to examine whether at least one of the moderator variables included in the model is related to the outcomes, or put differently, whether the set of moderators included in the model actually accounts for any heterogeneity in the true outcomes.

The fitted values based on the model can be computed with $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ or equivalently with $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$, where

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W} \tag{11.9}$$

is the hat matrix, whose relevance will be discussed in more detail further below. Similarly, for any row vector $\boldsymbol{x}_i$ (with $\boldsymbol{x}_i$ not necessarily a row from $\boldsymbol{X}$), we can compute the corresponding predicted value with $\hat{\mu}_i = \boldsymbol{x}_i\hat{\boldsymbol{\beta}}$, with variance equal to $\text{Var}[\hat{\mu}_i] = \boldsymbol{x}_i\text{Var}[\hat{\boldsymbol{\beta}}]\boldsymbol{x}_i'$ and standard error $SE[\hat{\mu}_i] = \sqrt{\text{Var}[\hat{\mu}_i]}$. Therefore, an approximate 95% CI for the expected true outcome given vector $\boldsymbol{x}_i$ can be computed with $\hat{\mu}_i \pm 1.96SE[\hat{\mu}_i]$. Note that in the mixed-effects model, $\hat{\mu}_i$ denotes the predicted *average* outcome for a particular combination of moderator values. The true outcome for a *particular* study can still differ from $\hat{\mu}_i$ due to residual heterogeneity.

The other three models can be considered to be special cases of the mixed-effects model. First, fixing $\tau^2 = 0$ yields the meta-regression model. The model coefficients and corresponding variance-covariance matrix are then estimated with (11.6) and (11.7) with $\boldsymbol{W} = \boldsymbol{V}^{-1}$. All other equations work accordingly, although notationally, it is now more appropriate to denote a predicted value with $\hat{\theta}_i$ (i.e., in the absence of residual heterogeneity, there is no longer a distinction between the predicted average outcome and the predicted outcome for a single study).

Second, the random-effects model results when $\boldsymbol{X}$ is only a column vector with the constant 1, in which case $\hat{\beta}_0 \equiv \hat{\mu} = \sum w_i y_i / \sum w_i$ with corresponding standard error $SE[\hat{\mu}] = \sqrt{1/\sum w_i}$, where $w_i = 1/(s_i^2 + \hat{\tau}^2)$. Note that (11.5) then simplifies to the well-known DerSimonian–Laird estimator of τ^2 in the random-effects model (DerSimonian and Laird, 1986).

Finally, when fixing $\tau^2 = 0$ and $\boldsymbol{X}$ is only a column vector with the constant 1, then we obtain the common-effect model, where $\hat{\beta}_0 \equiv \hat{\theta} = \sum w_i y_i / \sum w_i$ with corresponding standard error $SE[\hat{\theta}] = \sqrt{1/\sum w_i}$ as before, but now with $w_i = 1/s_i^2$.

An additional statistic that is often reported is the I^2 statistic (Higgins and Thompson, 2002). It is given by

$$I^2 = 100\% \times \left(\frac{\hat{\tau}^2}{\hat{\tau}^2 + \tilde{s}^2} \right), \tag{11.10}$$

where $\hat{\tau}^2$ is the estimate of τ^2 from the random-effects model and

$$\tilde{s}^2 = \frac{k-1}{\sum w_i - \sum w_i^2 / \sum w_i} \tag{11.11}$$

is a way of quantifying the "typical" sampling variance across the k studies, which is computed with $w_i = 1/s_i^2$. The I^2 statistic estimates what percentage of the total variability (which is composed of heterogeneity plus sampling variability) can be attributed to heterogeneity among the true outcomes. Also, for the mixed-effects model, a pseudo R^2-type measure can be computed with

$$R^2 = 100\% \times \left(\frac{\hat{\tau}_{RE}^2 - \hat{\tau}_{ME}^2}{\hat{\tau}_{RE}^2} \right), \tag{11.12}$$

where $\hat{\tau}_{RE}^2$ and $\hat{\tau}_{ME}^2$ are the estimates of τ^2 in the random- and mixed-effects models, respectively (López-López et al., 2014; Raudenbush, 2009). The R^2 statistic estimates the proportional reduction in the amount of heterogeneity when including moderators in a random-effects model, or put differently, what percentage of the total heterogeneity can be accounted for by the moderators included in the mixed-effects meta-regression model.

11.3 Example Dataset

For didactic purposes, a meta-analytic dataset was constructed that will be used to illustrate the application of the aforementioned models and the methods to be discussed in more detail below. The dataset, given in Table 11.1, can be thought of as a set of $k=20$ randomized controlled trials (e.g., patients receiving a medication versus a placebo) where a dichotomous response variable of interest was measured within the individual studies (e.g., remission versus persistence of symptoms). Columns n_i^T and n_i^C denote the total number of patients in the treatment and control group, respectively, while columns x_i^T and x_i^C denote the number of patients within the respective groups that experienced the outcome of interest (e.g., remission). The outcome measure to be used for the meta-analysis will be the log odds ratio.

The observed log odds ratios can be computed with

$$y_i = \ln\left[\frac{(x_i^T + 0.5)/(m_i^T + 0.5)}{(x_i^C + 0.5)/(m_i^C + 0.5)} \right], \tag{11.13}$$

where $m_i^T = n_i^T - x_i^T$ and $m_i^C = n_i^C - x_i^C$, and are given in the corresponding column in the table. A positive value for y_i therefore indicates higher odds of remission in the treatment compared with the control group. The sampling variances of the log odds ratios, given in the adjacent column, were computed with

$$s_i^2 = \frac{1}{x_i^T + 0.5} + \frac{1}{m_i^T + 0.5} + \frac{1}{x_i^C + 0.5} + \frac{1}{m_i^C + 0.5}. \tag{11.14}$$

TABLE 11.1

Illustrative Data for a Meta-Analysis of 20 Trials

Study	n_i^T	n_i^C	x_i^T	x_i^C	y_i	s_i^2	Dose
1	66	59	42	24	0.922	0.133	100
2	59	65	42	34	0.796	0.141	200
3	253	257	96	32	1.447	0.052	250
4	137	144	51	44	0.296	0.063	125
5	327	326	47	39	0.209	0.053	50
6	584	588	38	87	−0.907	0.041	25
7	526	532	390	323	0.617	0.018	125
8	28	30	10	3	1.495	0.471	125
9	191	201	165	126	1.316	0.065	125
10	86	94	58	39	1.059	0.096	150
11	229	221	72	60	0.206	0.043	100
12	153	144	79	56	0.514	0.055	150
13	93	95	48	35	0.597	0.087	200
14	40	40	8	4	0.752	0.398	25
15	85	88	44	21	1.214	0.108	175
16	100	107	10	13	−0.208	0.191	25
17	72	64	11	9	0.088	0.226	25
18	80	74	47	23	1.134	0.113	200
19	191	195	144	116	0.730	0.049	100
20	85	85	48	49	−0.047	0.095	75

The 0.5 term in the equations above serves two purposes. First, it reduces the bias in y_i as an estimate of θ_i, the true log odds ratio in the ith study (Walter, 1985; Walter and Cook, 1991). A second, more practical reason for the addition of the 0.5 term is that it allows the computation of the log odds ratio (and its corresponding sampling variance) even in studies where one of the 2×2 table cells (i.e., x_i^T, m_i^T, x_i^C, m_i^C) is equal to zero (although not applicable here, this issue will become relevant further below).

The last column in Table 11.1 reflects the dosage of the medication provided to patients in the treatment group (e.g., in milligrams per day), which we can envision as a potentially relevant moderator variable in this context. We will examine this moderator further with (mixed-effects) meta-regression models. In addition, common- and random-effects models were fitted to the data. Results for the various models are given in Table 11.2.

Under the assumption that the true outcomes are homogeneous, the common-effect model yields an estimate of the true log odds ratio equal to $\hat{\theta} = 0.529$ (95% CI: 0.415 to 0.644). Since the estimate is positive and the CI excludes the value 0, this suggests that the treatment significantly increases the odds of remission. Similarly, the random-effects model yields an estimate of $\hat{\mu} = 0.587$ (95% CI: 0.302 to 0.871), but the value should now be interpreted as the estimated *average* true effect of the treatment. In fact, we estimate that $I^2 = 81.70\%$ of the total variability can be attributed to heterogeneity, so it seems implausible that the true outcomes are homogeneous.

We can try to explain some of the heterogeneity with the dosage moderator. The meta-regression model suggests a significant relationship between dosage and the treatment effect, with the true log odds ratio increasing by $\hat{\beta}_1 = 0.008$ points for each additional

TABLE 11.2

Results for the Common-Effect, Random-Effects, Meta-Regression, and Mixed-Effects Models When Applied to the Data in Table 11.1 (Standard Error of the Estimates Are Given in Parentheses)

Common-effect model	Random-effects model	Meta-regression model	Mixed-effects model
$\hat{\theta} = 0.529$	$\hat{\mu} = 0.587$	$\hat{\beta}_0 = -0.423$	$\hat{\beta}_0 = -0.304$
(0.0586)	(0.1451)	(0.1333)	(0.2061)
$z = 9.03$	$z = 4.04$	$z_0 = -3.18$	$z_0 = -1.48$
		$\hat{\beta}_1 = 0.008$	$\hat{\beta}_1 = 0.007$
		(0.0010)	(0.0015)
		$z_1 = 7.96$	$z_1 = 4.76$
	$\hat{\tau}^2 = 0.317$		$\hat{\tau}^2 = 0.090$

milligram ($z_1 = 7.96$, so we can reject H_0: $\beta_1 = 0$). The mixed-effects model leads to a similar conclusion, except that $\hat{\beta}_1 = 0.007$ now reflects how the *average* true effect changes as a function of the treatment dosage (with $z_1 = 4.76$, the relationship is still significant). Both the random- and mixed-effects models indicate the presence of (residual) heterogeneity, although the dosage moderator appears to account for a substantial amount thereof. The pseudo R^2 statistic indicates that about $100\% \times (0.317 - 0.090) / 0.317 = 72\%$ of the total amount of heterogeneity is accounted for by the dosage moderator alone.

Results for the models are also illustrated graphically in Figure 11.1. The left-hand side of the figure shows a forest plot of the observed outcomes for the individual studies (with approximate 95% CI bounds for θ_i given by $y_i \pm 1.96 s_i$) and the results from the common-effect (CE) and random-effects (RE) models indicated at the bottom of the figure in terms of polygon shapes (with the center corresponding to the estimate and the ends corresponding to the 95% CI bounds). The right-hand side of Figure 11.1 shows a scatterplot of the dosage moderator on the x-axis versus the observed outcomes on the y-axis with the area of the points drawn proportional to the inverse sampling variances (i.e., $1/s_i^2$).

11.4 Checking Model Assumptions

The models described in Section 11.2 could be considered the main "workhorses" in meta-analytic applications. However, the models make various assumptions, which may be violated in practice. One of the crucial assumptions concerns the form of the sampling distribution of the chosen effect size or outcome measure. In particular, all of the models assume that $y_i \mid \theta_i \sim N(\theta_i, s_i^2)$, that is, conditional on θ_i, each observed outcome y_i is assumed to be drawn from a sampling distribution that is normal with expected value θ_i and variance s_i^2, with s_i^2 assumed to be a known and fixed quantity. We will now examine these assumptions in more detail and consider their plausibility.

11.4.1 Normal Sampling Distributions

Let us first examine the normality assumption in more detail. To begin with, it is important to emphasize that this assumption does not pertain to the collection of observed

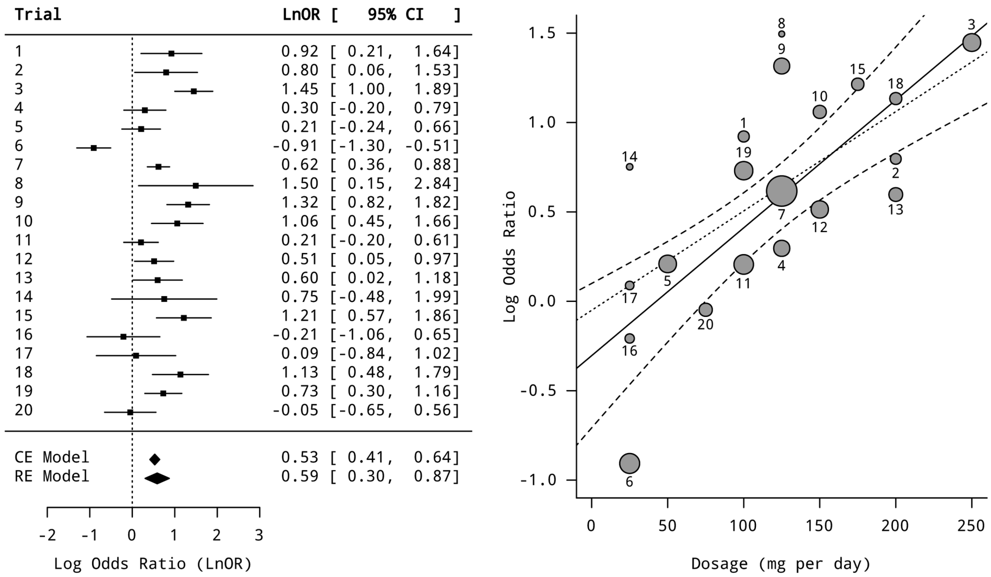


FIGURE 11.1
Forest plot showing the results of the CE and RE models and a scatterplot of the dosage moderator against the log odds ratios for the illustrative data in Table 11.1. The solid line in the scatterplot corresponds to the regression line (with intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$) for the mixed-effects model, with the dashed lines indicating (pointwise) 95% CIs around the predicted values. For reasons to be discussed further, the dotted line shows the regression line from the mixed-effects model when omitting study 6 from the dataset.

outcomes included in a meta-analysis (i.e., we do not assume that the observed y_i values themselves are normally distributed), but the theoretical distribution that would arise if a certain study were to be repeated a large number of times under identical circumstances. We shall construct this distribution now for one of the studies included in the illustrative dataset, again focusing on the log odds ratio as the chosen outcome measure.

Consider study 14, which included $n_i^T = n_i^C = 40$ patients in each group. Now, in order to construct the sampling distribution of the log odds ratio for this study, we would have to know π_i^T and π_i^C, the true probabilities of remission for patients in the treatment and control group, respectively. Obviously, these values are unknown to us, but we can use the observed proportions from this study (i.e., x_i^T / n_i^T and x_i^C / n_i^C) as an indication of the true probabilities and, for illustration purposes, simply assume that $\pi_i^T = .20$ and $\pi_i^C = .10$. Therefore, the true log odds ratio for this study would be equal to $\theta_i = \ln\left[(.20/(1-.20))/(.10/(1-.10))\right] = 0.811$. By taking random draws from the binomial distributions $B(n_i^T = 40, \pi_i^T = .20)$ and $B(n_i^C = 40, \pi_i^C = .10)$, we can then easily generate values for x_i^T and x_i^C, which we can use to calculate the log odds ratio as given by (11.13). By repeating this process a large number of times under identical circumstances (i.e., keeping n_i^T, n_i^C, π_i^T, and π_i^C the same), we can generate the sampling distribution of the log odds ratio under the described scenario. Use of the 0.5 adjustment term as described earlier guarantees that we can compute the log odds ratio in every iteration of such a simulation.

Figure 11.2 shows the shape of the distribution generated in this manner (after 10^7 iterations) in terms of a histogram and a normal quantile-quantile (Q-Q) plot. The solid line

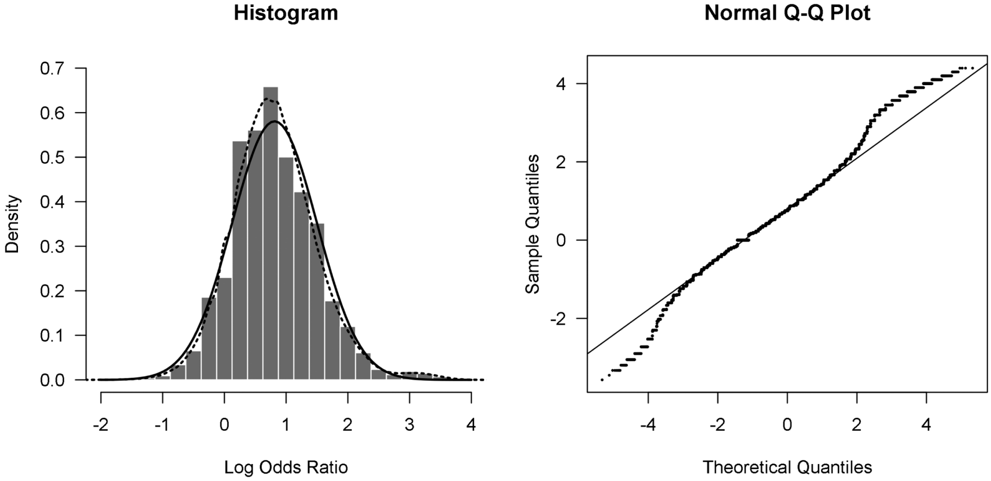


FIGURE 11.2
Histogram and normal Q-Q plot of the sampling distribution of the log odds ratio for a study with $n_i^T = n_i^C = 40$ patients in the treatment and control group and true probabilities of $\pi_i^T = .20$ and $\pi_i^C = .10$, respectively.

superimposed on the histogram corresponds to a normal distribution with mean and standard deviation equal to that of the simulated log odds ratios, while the dotted line represents a kernel density estimate of the underlying distribution. The log odds ratios ranged from −3.570 to 4.394 with a mean of 0.812 (SD = 0.687). Therefore, under the simulated conditions, (11.13) provides an essentially unbiased estimate of $\theta_i = 0.811$. However, the shape of the sampling distribution deviates slightly from that of a normal distribution, especially in the tails. Although the departure from normality may be negligible in this case, the example does illustrate that we cannot assume that the normality assumption is automatically fulfilled.

In fact, the normality assumption is only approximately true for most of the commonly used outcome measures that are applied in the meta-analytic context. Certainly, for measures that are bounded (e.g., the raw correlation coefficient or the risk difference), the assumption cannot be true in general, but it may still hold as a rough approximation. At the same time, for unbounded measures (such as the log odds ratio or the standardized mean difference), it may be quite inaccurate under certain circumstances. In particular, when sample sizes within studies are small and/or when the underlying parameters are very large or small (e.g., π_i^T and π_i^C are close to 0 or 1 for measures such as the log odds/risk ratio or the true standardized mean difference is far from 0), the assumption can break down altogether. Also, due to the discrete nature of the counts on which they are based, measures such as the log odds ratio can only generate a finite number of unique values. For example, in the simulation above, only 310 unique values of the log odds ratio were observed across the 10^7 iterations.* Strictly speaking, the sampling distribution cannot be normal then, but we can still consider the values to be discretized observations from an underlying normal distribution.

* Technically, $41^2 = 1681$ combinations of x_i^T and x_i^C are possible, although some of them (e.g., those where $x_i^T = x_i^C$) lead to the same value of y_i. Simple enumeration of all possibilities shows that there are only 1083 unique values of y_i that could be observed, but many of them are so unlikely to happen that they do not occur even once across such a large number of iterations.

11.4.2 Unbiased Estimates

The second assumption implied by the models is that y_i is an unbiased estimate of θ_i. In the simulation above, the log odds ratio computed with (11.13) was found to be essentially unbiased in the given scenario. Moreover, prior work has shown that (11.13) yields an approximately unbiased estimator across a wide range of conditions (e.g., Walter, 1985; Walter and Cook, 1991). Interestingly, this work also indicates that the 0.5 adjustment used in the equation should always be applied, not just in studies where the presence of a zero cell in the 2×2 table would necessitate its use.

Similarly, the bias in estimators for other outcome measures has also been examined, leading to known bias corrections for the standardized mean difference (Hedges, 1981), the raw correlation coefficient (Olkin and Pratt, 1958), and Fisher's r-to-z transformed correlation coefficient (Hotelling, 1953). On the other hand, the risk difference and the raw mean difference are unbiased by construction. Generally, for estimators that are biased, the amount of bias usually diminishes quickly as the within-study sample size increases.

11.4.3 Known Sampling Variances

Next, we will consider the assumption that the sampling variances of the observed outcomes are known and fixed quantities. This assumption is often not exactly true for two reasons. First, for many outcome measures, commonly used equations to compute the sampling variances are based on derivations that describe the asymptotic (i.e., large-sample) properties of the outcome measures. Hence, in finite samples, the equations may not be accurate. Second, the equations often depend on unknown parameters. To compute the sampling variances, these unknown parameters must be replaced by corresponding estimates, which introduces error into the values.

Consider again the log odds ratio. The asymptotic sampling variance of y_i can be shown to be equal to

$$s_i^2 \overset{\infty}{=} \frac{1}{n_i^T \pi_i^T} + \frac{1}{n_i^T (1-\pi_i^T)} + \frac{1}{n_i^C \pi_i^C} + \frac{1}{n_i^C (1-\pi_i^C)}, \tag{11.15}$$

where π_i^T and π_i^C are the true probabilities of the event of interest occurring in the treatment and control group, respectively. Note that this equation is technically only correct when n_i^T and n_i^C are sufficiently large, which raises the question how large the groups need to be for the equation to be accurate. Moreover, since π_i^T and π_i^C are unknown parameters, the observed probabilities $p_i^T = x_i^T / n_i^T$ and $p_i^C = x_i^C / n_i^C$ are typically substituted, leading to equation (11.14) (with the addition of the 0.5 term to make the computation of s_i^2 possible under all circumstances). The resulting s_i^2 values are therefore not truly fixed and known quantities but are estimates themselves.

We can illustrate this again using the simulated data from the previous section. Given the assumed values of $\pi_i^T = 0.20$ and $\pi_i^C = 0.10$, we can compute the (large-sample) variance with (11.15), which is equal to $s_i^2 \overset{\infty}{=} 0.434$ in this case. However, taking the variance of the simulated y_i values yields $\text{Var}[y_i] = 0.472$, which we can consider to be the true sampling variance of the log odds ratio under the given scenario (the simulation error is negligible due to the very large number of values generated). Therefore, the actual sampling variance is about 9% larger than what we obtain with (11.15), so some inaccuracy is introduced by basing our computations on an equation that describes the large-sample properties of the log odds ratio.

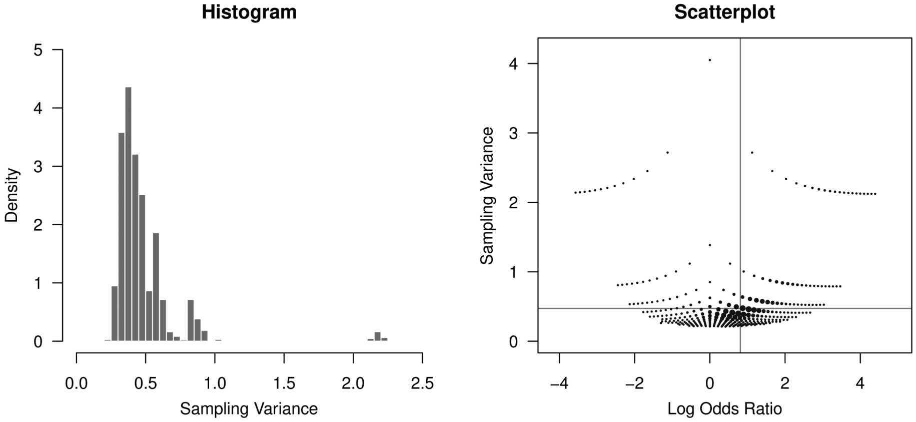


FIGURE 11.3
Distribution of the observed sampling variances of the log odds ratio for a study with $n_i^T = n_i^C = 40$ patients in the treatment and control group and true probabilities of $\pi_i^T = .20$ and $\pi_i^C = .10$, respectively.

However, neither of these two values would be available in practice. Instead, one would compute the sampling variance with (11.14) by plugging in the x_i^T and x_i^C values observed in a given sample. As a result, the calculated s_i^2 value may under- or overestimate the true sampling variance, sometimes to a considerable degree. The simulation above allows us to explore the range of values one could obtain under the stated scenario. Across all iterations, the s_i^2 values ranged from 0.210 to 4.049, although very large values were rare, as can be seen in the histogram of the observed s_i^2 values in Figure 11.3. Values above 2.5 were only seen in 316 out of the 10^7 iterations (and hence, the x-axis in the histogram was restricted to a range of 0 to 2.5) and usually fell below 1 in 98% of the cases. However, even then, the observed s_i^2 value could still provide a rather inaccurate estimate of the true sampling variance (i.e., 0.472 as noted above). At least somewhat reassuring is the finding that the average of the observed values was 0.482, so (11.14) provides an approximately unbiased estimate of the true sampling variance in this scenario.

Interestingly, if we plot the observed log odds ratios against the estimated sampling variances, a peculiar pattern emerges as can be seen in the right-hand side of Figure 11.3 (the vertical and horizontal lines indicate the true log odds ratio and true sampling variance, respectively). The plot illustrates the discrete nature of possible values that can actually arise (and therefore, the size of the points was drawn proportional to the number of observed values for a particular $\langle y_i, s_i^2 \rangle$ combination). Clearly, the two sets of values are related to each other (in this example, the correlation between the y_i and s_i^2 values is 0.57), a fact that has been noted before (Berkey et al., 1995; Rücker et al., 2008). In summary then, it seems inappropriate to assume that the sampling variances of the observed outcomes are really fixed and known quantities, a point that has been emphasized in Chapter 5.

In general, the problem that the sampling variances are inaccurately estimated (and hence, falsely treated as fixed and known quantities) tends to become more severe in smaller samples and when the underlying parameters are very large or small. The same issue applies, to a larger or lesser extent, to many other outcome measures used in meta-analysis, such as the risk difference, the log risk ratio, the raw and the standardized mean difference, and the raw correlation coefficient. An exception to this are measures based on a variance stabilizing transformation, most notably Fisher's r-to-z transformed correlation coefficient (Fisher, 1921) and the arcsine (and square root) transformed risk difference (Rücker et al., 2009), although even here we need to be precise: The variance stabilizing transformation eliminates the unknown parameters from the equation used to compute the sampling variance, but the resulting equation is still in principle a large-sample approximation. However, for both the r-to-z correlation coefficient and the arcsine transformed risk difference, the approximation is surprisingly accurate even for relatively small studies.

11.4.4 Uncorrelated Errors and Random Effects

One additional assumption pertains to the random- and mixed-effects models. Recall that these models contain two random effects, namely u_i, which represents a deviation of the study-specific true outcome from the average true outcome of all studies (or from the average true outcome for those studies that share a particular combination of moderator variables) and e_i, the sampling error that represents a deviation of the observed outcome from the study-specific true outcome. The models assume that u_i and e_i are independent.

Once again, we will examine whether this assumption is appropriate for the log odds ratio by means of a simulation study. Here, we proceed as follows. As before, assume we are dealing with a study with $n_i^T = n_i^C = 40$ patients in each group and that $\pi_i^C = .10$. For a given true log odds ratio, θ_i, we can compute the implied value of the treatment group probability with $\pi_i^T = \pi_i^C \exp(\theta_i) / (1 - \pi_i^C + \pi_i^C \exp(\theta_i))$. For example, for $\theta_i = 0.811$, this yields $\pi_i^T = .20$ as expected. However, instead of fixing θ_i, we now let θ_i follow a normal distribution, as assumed by the random-effects model. Specifically, we draw u_i from $N(0, \tau^2 = .20)$ and then compute $\theta_i = \mu + u_i$ with $\mu = 0.811$. Next, we compute π_i^T and then draw x_i^T and x_i^C values from the respective binomial distributions. Finally, we can compute y_i as described earlier and then $e_i = y_i - \theta_i$ yields the sampling error for one iteration.

Repeating this process 10^6 times yields pairs of $\langle u_i, e_i \rangle$ values, whose bivariate distribution we can now examine, for example, by means of the scatterplot shown on the left-hand side of Figure 11.4 (due to the large number of points drawn, alpha blending* was used to make differ-

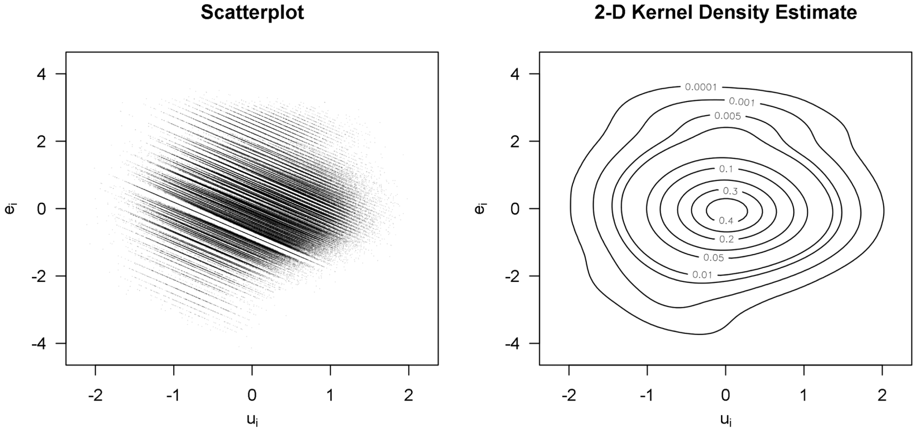


FIGURE 11.4
Scatterplot and 2D kernel density estimate of the bivariate distribution of u_i and e_i.

* In alpha blending, each point is drawn with a certain degree of transparency. Multiple points drawn on top of each other then blend together to create increasing darker shades of the plotting color. As a result, the color intensity indicates the density of points within a particular region of the plot.

ences in the density of the distribution more apparent). The right-hand side shows a 2D kernel density estimate of the bivariate distribution. The patterns in the scatterplot are again a result of the discrete nature of the distribution of y_i values that can arise. However, the kernel density estimate suggests a unimodal, roughly symmetric distribution with no apparent relationship between the u_i and e_i values (the correlation is zero to three decimals). Although this does not demonstrate independence (zero correlation only implies independence if $\langle u_i, e_i \rangle$ would follow a bivariate normal distribution), it does provide some support for the independence assumption in this scenario.

To what extent the independence assumption holds in other scenarios, and more generally, for other outcome measures besides the log odds ratio, has not been examined in detail in the literature. For the raw mean difference, u_i and e_i are independent by construction. On the other hand, for measures where the shape of the sampling distribution depends on the size of the underlying parameter estimated (e.g., the raw correlation coefficient), it is easy to reason that u_i and e_i will not form a bivariate normal distribution with zero correlation, especially in small samples. However, further research is needed to determine if or to what degree u_i and e_i are dependent in such cases.

11.4.5 Implications and Some General Remarks

The log odds ratio was given particular attention in this section, since this was the outcome measure of choice in the example meta-analysis. However, analogous considerations apply to other outcome measures for dichotomous response variables (e.g., the log risk ratio and the risk difference), outcome measures for continuous response variables (e.g., the raw mean difference and the standardized mean difference), and outcome measures used to quantify the relationship between variables (e.g., the raw or Fisher's r-to-z transformed correlation coefficient).

In fact, a careful examination shows that there is not a single outcome measure used in meta-analysis that fulfills all of the assumptions described above exactly. Most of them actually fulfill very few of them! The raw mean difference could be argued to comes closest by construction, having an exactly normal sampling distribution, being unbiased, with uncorrelated errors and random effects, although its sampling variance must still be estimated using sample quantities. The risk difference is also unbiased and when appropriate bias corrections are applied, the same holds for the standardized mean difference (Hedges, 1981), the raw correlation coefficient (Olkin and Pratt, 1958), and Fisher's r-to-z correlation coefficient (Hotelling, 1953). Approximate unbiasedness can also be demonstrated by means of simulation studies for some of the other measures (possibly requiring some adjustments to the way the measure is typically computed), although this cannot hold at the same level of generality as when a measure can be shown to be unbiased with an analytic proof. Similarly, although no other measure besides the raw mean difference has an exactly normal sampling distribution, simulation studies can be used to examine to what extent and under what circumstances this assumption is at least approximately fulfilled for other measures. Finally, while it is demonstrably false to assume known sampling variances for any of the measures, those that involve a variance stabilizing transformation at least fulfill this assumption approximately under most circumstances.

At the same time, it can be shown for essentially all measures that the underlying estimators are consistent, asymptotically unbiased, and that their sampling distribution approaches the shape of a normal distribution as the within-study sample size increases (note that for measures that reflect a contrast between two independent samples—such as the log odds ratio or the standardized mean difference—this requires that both n_i^T and n_i^C increase with similar rates).

The discussion above then raises two important questions. First, under what circumstances do the assumptions break down to such an extent that we should no longer apply these methods and models? And second, what are the consequences when one or more of the assumptions are violated? Unfortunately, there are no simple answers to these questions, because the answers depend on various factors, including the outcome measure under consideration, the size of the studies, and whether underlying parameters may be extreme in some sense and/or close to their natural bounds. However, for all measures, there are circumstances where serious concerns should be raised about their use, whether due to bias, non-normality, or other violations. As a result, point estimates may not be trustworthy and/or tests and CIs may not have their nominal properties (i.e., the actual type I error rate of hypothesis tests and the actual coverage of CIs may deviate from the chosen significance/confidence level).

There is a large amount of literature that has examined the statistical properties of the methods and models described above for a wide variety of outcome measures and conditions (e.g., Berkey et al., 1995; Field, 2001; Friedrich et al., 2008; Hauck et al., 1982; Hedges, 1982a, 1982b, 1982c; Sánchez-Meca and Marín-Martínez, 2001 just to give a small selection). While it would require many more pages to discuss the details, the general conclusion is that for many of the measures, the methods perform adequately, even when the assumptions discussed above are not exactly fulfilled. However, this should not be taken as an *a priori* truth. Moreover, most studies have examined the properties of the methods under circumstances where the assumptions underlying the *construction* of the outcome measures are exactly fulfilled (e.g., for the standardized mean difference, the data within the two groups was simulated from normal distributions). More work is still needed to examine the robustness of the methods when such underlying assumptions are also violated.

11.5 Checking Model Fit

The assumptions discussed in the previous section relate to the statistical properties of the outcome measure chosen for the meta-analysis and should be carefully considered before we even start fitting models of the type described in Section 11.2. Once we are satisfied that these assumptions are (at least approximately) fulfilled, the actual analysis can then proceed by fitting one or more models to the data at hand.

It is important to emphasize that a statistical model in essence represents an assumption about the underlying data generating process. For example, when we fit the common-effect model to the data from the illustrative example, we implicitly assume that the data within the individual studies arose in such a way that the true treatment effect (as measured in terms of the log odds ratio) is constant across the trials, irrespective of any differences in the way the studies were designed, conducted, how the treatment was implemented/administered, how the dependent variable was measured, or any relevant patient characteristics. This is unlikely to be exactly true but may hold as a rough approximation in some applications (e.g., when a series of trials is conducted using identical methods in very similar patient populations).

Similarly, if we reject the assumption that the true outcomes are homogeneous across trials and decide to fit a meta-regression model with medication dose as a moderator, we implicitly assume that any differences in treatment effectiveness across trials can be accounted for with this single explanatory variable. Again, this is likely to be a great oversimplification of a much more complex reality, but it may serve as an adequate approximation. To check whether these assumptions are actually appropriate, we can apply tests for model misspecification, which are covered next.

11.5.1 Testing for Model Misspecification

The homogeneity assumption underlying the common-effect model can be tested by means of the Q-test. Under the assumptions of the model, the test statistic

$$Q = \sum w_i (y_i - \hat{\theta})^2 \tag{11.16}$$

follows a chi-square distribution with $k-1$ degrees of freedom, where $w_i = 1/s_i^2$ and $\hat{\theta}$ is the estimate of θ under the common-effect model. However, when the observed outcomes fluctuate more around $\hat{\theta}$ than expected based on their sampling variances alone (which should be the only source of variance affecting the y_i values under the common-effect model), the Q-statistic increases. Therefore, a large Q-statistic should lead us to question the correctness of the assumed model.

Similarly, the meta-regression model (11.3) assumes that all of the heterogeneity in the true outcomes can be accounted for with the moderator variables included in the model, or analogously, that the model $\theta_i = \boldsymbol{x}_i \boldsymbol{\beta}$ is correctly specified. Again, this assumption can be tested, using a generalization of the Q-test to the meta-regression model. Here, the test statistic is computed with

$$Q_E = \sum w_i (y_i - \hat{\theta}_i)^2, \tag{11.17}$$

where $\hat{\theta}_i = \boldsymbol{x}_i \hat{\boldsymbol{\beta}}$ is the fitted value for the ith study from the meta-regression model (see Section 11.2.5) and $w_i = 1/s_i^2$ as before (note that this is the same Q_E statistic that is also involved in the estimator for τ^2 given by equation 11.5). If the assumed meta-regression model correctly describes the underlying data generating process, then the Q_E statistic follows a chi-square distribution with $k-q-1$ degrees of freedom. Again, as the degree of mismatch between the observed outcomes and the fitted values under the assumed model increases (i.e., more so than would be expected based on sampling variability alone), the Q_E statistic grows larger.

Although not frequently described in this manner, we can therefore consider (11.16) and (11.17) to be tests of model misspecification (Hedges, 1992). Rejection of the null hypothesis (that the model is adequately specified) should then be taken as an indication that the assumed model does not provide an adequate approximation to the underlying data generating process. At the same time, non-rejection must be cautiously interpreted, as the tests may lack power especially when the number of studies is small.

For the example data, we find $Q = 103.81$, a value so extreme that the chances of it (or an even larger value) occurring under a chi-square distribution with 19 degrees of freedom are extremely small (i.e., $p < 0.0001$). Consequently, we would reject the common-effect model as a plausible approximation. The same applies to the meta-regression model, for which we find $Q_E = 40.44$, also a rather unlikely occurrence under a chi-square distribution with 18 degrees of freedom (i.e., $\mathrm{p} = 0.002$).

There are no analogous versions of these types of tests for the random- and mixed-effects models. However, we can resort to a different approach by examining the residuals (and standardized versions thereof) computed from the fitted model, which may reveal a mismatch between particular data points and the assumed model.

11.5.2 Residuals and Standardized Versions Thereof

The raw residual for the ith study is $y_i - \hat{\mu}_i$ (in the common- and random-effects models, $\hat{\mu}_i$ is simply $\hat{\theta}$ and $\hat{\mu}$, respectively). The raw residuals should scatter randomly around the fitted values, but are not very useful for diagnostic purposes, as they do not account for differences in the sampling variances across studies. Moreover, in the random-effects and mixed-effects models, (residual) heterogeneity represents an additional source of variability, which needs to be taken into consideration. Instead, we can compute Pearson (or semi-standardized) residuals, which are given by

$$r_i = \sqrt{w_i}(y_i - \hat{\mu}_i), \tag{11.18}$$

where $w_i = 1/s_i^2$ in the common-effect and meta-regression models and $w_i = 1/(s_i^2 + \hat{\tau}^2)$ in the random- and mixed-effects models. As can be seen from (11.16) and (11.17), for the common-effect and meta-regression models (where $\hat{\mu}_i = \hat{\theta}$ and $\hat{\mu}_i = \hat{\theta}_i$, respectively), the Q and Q_E statistics are just the sum of the squared Pearson residuals, or put differently, r_i^2 is the contribution of the ith study to these statistics.

However, Pearson residuals are not properly standardized (in the sense of having unit variances), as they do not account for the imprecision in the $\hat{\mu}_i$ values. In fact, it follows from the results laid out in Section 11.2.5 that the variance of the raw residual in the ith study can be estimated with $(1-h_i)(s_i^2 + \hat{\tau}^2)$, where h_i is the ith diagonal element from the hat matrix $\boldsymbol{H}$ (for the common-effect and meta-regression models, $\hat{\tau}^2 = 0$ by definition). Therefore, the (internally) standardized residuals can be computed with

$$\tilde{r}_i = \frac{y_i - \hat{\mu}_i}{\sqrt{(1-h_i)(s_i^2 + \hat{\tau}^2)}}, \tag{11.19}$$

which have approximately unit variances (imprecision is introduced due to s_i^2 and $\hat{\tau}^2$ being estimates themselves).

We will consider one other type of residual, which is related to the deletion diagnostics to be discussed in more detail further below. Here, we compute the residual of a study based on a model that excludes the study during the model fitting process. Therefore, we first delete the ith study from the dataset and then fit the model of choice using the remaining $k-1$ studies. When fitting a random- or mixed-effects models, let $\hat{\tau}^2_{(-i)}$ denote the estimate of τ^2 from this model (i.e., the $(-i)$ part in this and other subscripts will be used to indicate that the value was computed from the fitted model that excluded the ith study from the model fitting). Next, we compute the predicted value for the study that was deleted, which we denote by $\hat{\mu}_{i(-i)}$. Furthermore, let $\text{Var}[\hat{\mu}_{i(-i)}]$ denote the corresponding variance of the predicted value. Then we define the "deleted residual" as $r_{i(-i)} = y_i - \hat{\mu}_{i(-i)}$. Finally,

$$t_i = \frac{y_i - \hat{\mu}_{i(-i)}}{\sqrt{s_i^2 + \hat{\tau}^2_{(-i)} + \text{Var}[\hat{\mu}_{i(-i)}]}} \tag{11.20}$$

yields the standardized deleted residual (or externally standardized residual), which again has approximately unit variance (Viechtbauer and Cheung, 2010). Note that $\hat{\tau}^2_{(-i)} = 0$ for the common-effect and meta-regression models by definition. Also, the notation can be further simplified depending on the model (i.e., $\hat{\mu}_{i(-i)} = \hat{\theta}_{i(-i)}$ for the meta-regression model, $\hat{\mu}_{i(-i)} = \hat{\mu}_{(-i)}$ for the random-effects model, and $\hat{\mu}_{i(-i)} = \hat{\theta}_{(-i)}$ for the common-effect model).

11.5.3 Checking for Outliers

Standardized (deleted) residuals are useful for detecting outliers, that is, studies that do not fit the assumed model. The advantage of the deleted residuals is that they are more sensitive to detecting outliers. In particular, if a study does not fit the assumed model, then this affects the results in two ways. First, the study will introduce additional heterogeneity into the data (i.e., $\hat{\tau}^2$ tends to increase), which will get subsumed into the standard deviation of the residuals, as shown in the denominator of (11.19). This will shrink the standardized residuals toward 0 to some degree, making it more difficult to detect the outlying study. In addition, the y_i value of an outlying study will pull $\hat{\mu}_i$ toward it, leading to a smaller raw residual, and hence, a smaller standardized residual. By first deleting a potentially outlying study from the dataset, both of these effects are eliminated, making the standardized deleted residual for the study a more sensitive indicator of whether the study fits the model or not.

Assuming that the fitted model is the correct one for the data at hand, the probability that a standardized (deleted) residual is larger than ±1.96 is approximately 5%. Therefore, while we may expect to observe one or maybe two large values in a set of $k = 20$ studies, the presence of a large number of such values would indicate that the assumed model does not represent a good approximation to the underlying data generating process. Based on the binomial distribution (with $\pi = 0.05$ and $k = 20$), the chances of observing one or more, two or more, and three or more standardized (deleted) residuals this large are 64%, 26%, and 8%, respectively.

For the example data, we find four standardized deleted residuals larger than ±1.96 for the common-effect model and three for the meta-regression model, again suggesting a mismatch between the data and these models. How do the random- and mixed-effects models fare? Figure 11.5 shows the standardized deleted residuals for these models. Only one and two values larger than ±1.96 are found in these models, respectively, so we do not have grounds to question the adequacy of these models in general terms. However, study 6 appears to be an outlier in both of these models, which is also recognizable from inspection of the plots in Figure 11.1. In the mixed-effects model, the standardized deleted residual of study 9 is also larger than 1.96, although just barely.

Interestingly, the scatterplot in Figure 11.1 shows that the estimate of study 8 is further away from the regression line than that of study 9 (which may suggest that study 8 is more of an outlier than study 9), but the standardized deleted residual for the latter is larger. This seeming contradiction can be explained by the fact that study 9 included almost seven times as many participants and hence has a much smaller sampling variance than study 8. Therefore, a deviation as large as the one for study 8 could be accounted for based on its sampling variability (plus residual heterogeneity and variance in the predicted value), while the residual of study 9 is more unusual given its smaller sampling variance. Also, Figure 11.1 suggests that study 14 has an unusually large outcome for a study with such

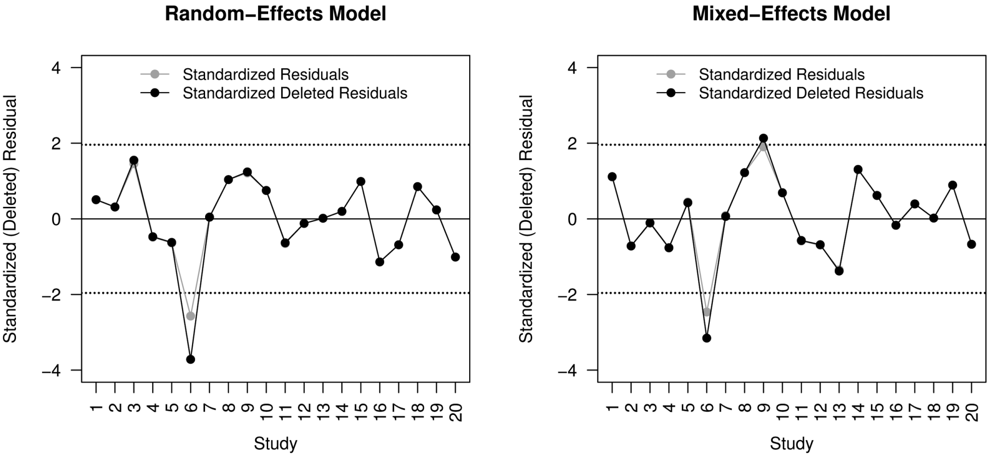


FIGURE 11.5
Standardized (deleted) residuals for the random- and mixed-effects models.

a low dosage, but its standardized deleted residual indicates that this deviation is not anomalous under the model. Hence, while forest and scatterplots can be useful visual aids for detecting outlying studies, they can also be deceptive.

Besides the standardized deleted residuals, Figure 11.5 also shows the regular standardized residuals computed with (11.19). While there is not much of a difference between these two types of residuals for most studies, we do see how the standardized deleted residuals are more sensitive to potential outliers, especially for studies 6 and 9. For example, for study 6, the deleted residual (i.e., the deviation of the outcome of the study from the dotted regression line in the scatterplot in Figure 11.1) is considerably larger than the regular residual (i.e., the deviation from the solid regression line). In addition, $\hat{\tau}^2_{(-6)} = 0.032$ is much smaller than the estimate of residual heterogeneity when all studies are included (i.e., $\hat{\tau}^2 = 0.090$), further leading to a more sensitive test.

11.5.4 Baujat and GOSH Plots

Several other graphical tools have been suggested in the literature for detecting outliers and sources of (residual) heterogeneity in meta-analytic data. We will now consider two of these devices and illustrate their use by applying them to the example dataset.

First, we will consider a type of plot suggested by Baujat and colleagues (Baujat et al., 2002). As originally described, the plot shows the contribution of each study to the *Q*-statistic on the x-axis versus the influence of each study on the overall estimate from a common-effect model on the y-axis. However, the idea underlying this type of plot can be easily generalized to random-effects models and models including moderators (with the common-effect model again forming a special case). In general, we plot the squared Pearson residual of each study (i.e., r_i^2) on the x-axis against

$$\frac{(\hat{\mu}_i - \hat{\mu}_{i(-i)})^2}{\mathrm{Var}[\hat{\mu}_{i(-i)}]}, \tag{11.21}$$

that is, the standardized squared difference between the predicted/fitted value for a study with and without the study included in the model fitting. Hence, a study whose observed

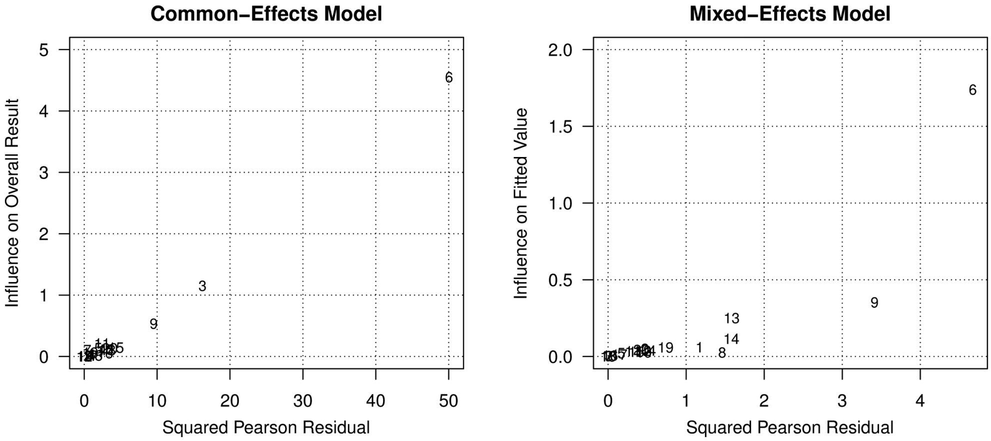


FIGURE 11.6
Baujat plots for the common-effect and mixed-effects models.

outcome deviates strongly from its predicted value based on the model will be located on the right-hand side of the plot. For the common-effect and the meta-regression models, these are the studies that contribute most to the Q (i.e., equation 11.16) and Q_E (i.e., equation 11.17) statistics. Furthermore, a study whose predicted value changes strongly depending on whether it is included or excluded from the dataset will be located on the top-hand side of the plot.

Figure 11.6 shows Baujat plots corresponding to the common-effect and mixed-effects models for the illustrative dataset. As we saw earlier, studies 6 and 9 show up again as apparent outliers in the context of the mixed-effects model. On the other hand, the plot for the common-effect model indicates that study 3 also contributes considerably to the overall amount of heterogeneity.

Another type of plot that is useful for detecting outliers and sources of heterogeneity is the so-called GOSH (graphical display of study heterogeneity) plot (Olkin et al., 2012). As originally described, the plot is constructed as follows. First, we fit the common-effect model to all possible subsets of size $1,\ldots,k$ of the k studies included in a meta-analysis. Therefore, at the one extreme, this will include k models that are each fitted to a single observed outcome (in which case $\hat{\theta} = y_i$), then the $\begin{pmatrix} k \\ 2 \end{pmatrix}$ models fitted to all pairwise combinations of two observed outcomes, and so on, until we get to the original model using all k outcomes. In total, there are then $\sum_{i=1}^{k} \begin{pmatrix} k \\ i \end{pmatrix} = 2^k - 1$ models that need to be fitted. We can then plot the model estimates obtained this way (e.g., as a histogram and/or using a kernel density estimate) to examine the resulting distribution. In a homogeneous set of studies, the distribution should be roughly symmetric, contiguous, and unimodal. On the other hand, when the distribution is multimodal, then this suggests the presence of heterogeneity, possibly due to the presence of outliers and/or distinct subgroupings of studies. Plotting the estimates against some measure of heterogeneity (e.g., I^2) computed within each subset can also help to reveal subclusters, which are indicative of heterogeneity.

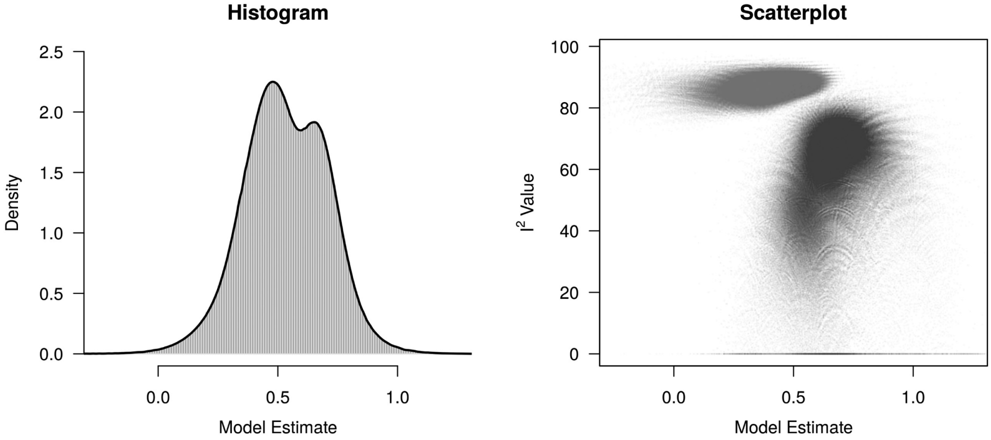


FIGURE 11.7
GOSH plot showing the distribution of estimates from the common-effect model based on all possible subsets and the bivariate distribution of the estimates and the corresponding I^2 values (results from subsets including study 6 are shown in light and dark gray otherwise).

For the illustrative dataset, a total of $2^{10} - 1 = 1{,}048{,}575$ subsets can be constructed. The left-hand side of Figure 11.7 shows a histogram of the model estimates when fitting the common-effect model to each of these subsets (with a kernel density estimate of the distribution superimposed). The bimodal shape of the distribution is a result of study 6, which has considerable impact on the model estimate depending on whether the study is included in a subset or not. The influence of this study becomes even more apparent in the plot on the right-hand side of Figure 11.7, which shows a scatterplot of the model estimates against the corresponding I^2 values (using alpha blending). Results from subsets including study 6 are shown in light and dark gray otherwise. Inclusion of study 6 in a subset not only tends to reduce the size of the model estimate, but also increases the percentage of variability that can be attributed to heterogeneity.

The idea underlying the GOSH plot can be generalized to other types of models (including models with moderator variables) by examining the distribution of all model coefficients across subsets, plotting them against each other, and against some measure of (residual) heterogeneity (e.g., $\hat{\tau}^2$). Note that for a model with q moderator variables and an intercept term, a subset must include at least $q+1$ studies for the model to be estimable. When fitting a mixed-effects model, at least $q+2$ studies must be included in a subset so that τ^2 can also be estimated. Even then, a model may not be estimable in certain subsets. For example, for the (mixed-effects) meta-regression model with the dosage moderator, the subset including studies 6, 14, 16, and 17 does not allow estimation of β_1, since all four studies were conducted at the same dosage level of 25 mg. Therefore, only subsets where all parameters are estimable can be used for creating the plot.

Figure 11.8 shows such a generalized GOSH plot for the mixed-effects model (a total of 1,048,353 models could be fitted). The figure shows the distribution of $\hat{\tau}^2$, $\hat{\beta}_0$, and $\hat{\beta}_1$, and all pairwise scatterplots. Again, subsets including study 6 are shown in light and dark gray otherwise. As expected (cf. the scatterplot in Figure 11.1), for subsets that do include study 6, the estimate of residual heterogeneity tends to be higher, the model intercept tends to be lower, and the slope tends to be steeper. As a result, the distributions are again bimodal.

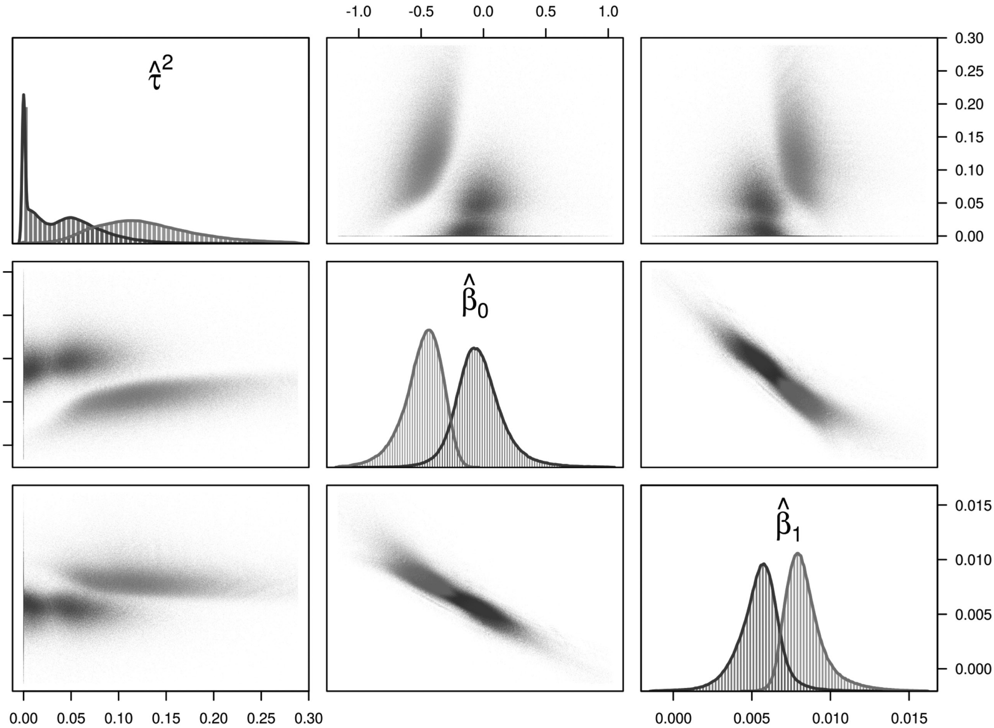


FIGURE 11.8

GOSH plot for the mixed-effects model showing the distribution of $\hat{\tau}^2$, $\hat{\beta}_0$, and $\hat{\beta}_1$ and all pairwise scatterplots based on all possible subsets (results from subsets including study 6 are shown in light and dark gray otherwise).

11.5.5 Testing for Lack of Linearity

For the illustrative dataset, we saw earlier that the meta-regression models suggest an increasing relationship between the treatment dosage and the (average) log odds ratio. However, when fitting meta-regression models that include continuous moderators, an aspect to consider is the linearity of the assumed relationship. Although the scatterplot in Figure 11.1 does not suggest any non-linearity of the relationship, several approaches can be used to examine the linearity assumption more systematically.

As in regular regression modeling (e.g., Kutner et al., 2004), one can examine a plot of each continuous moderator variable against the residuals from the (mixed-effects) meta-regression model to examine whether patterns are visible that may suggest potential non-linearity (e.g., a U- or an upside-down U-shape). For a model with a single continuous moderator variable, such a plot is not fundamentally different than just the scatterplot of the moderator variable against the observed outcomes, although it can be easier to detect patterns when the linear trend has been removed from the data by computing the residuals. Also, since the residuals are heteroscedastic (in part due to the heteroscedastic nature of the sampling variances), it can be useful to place the standardized (deleted) residuals on the y-axis (which should have roughly unit variance). The left-hand side of Figure 11.9 shows such a plot of the standardized residuals from the mixed-effects model. Again, we notice the outlier (study 6) in the lower left-hand corner of the plot, but otherwise no apparent curvature in the point cloud.

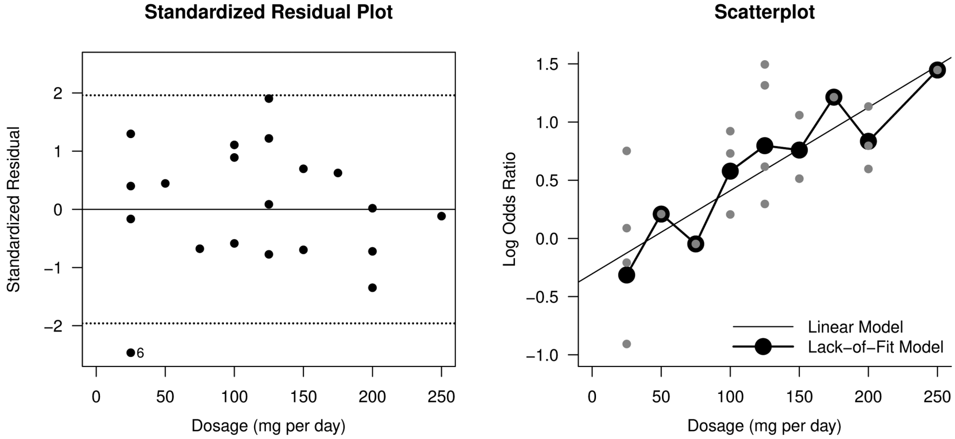


FIGURE 11.9
Plot of treatment dosage against the standardized residuals (from the mixed-effects model) and against the observed log odds ratios.

It is also possible to test more formally whether the relationship between dosage and the outcomes deviates from linearity. As a simple first approximation, one could consider fitting polynomial meta-regression models (e.g., adding the squared dosage as an additional moderator to the mixed-effects model). Doing so in the present case leads to a non-significant coefficient for the squared dosage term of the model ($z_2 = -1.26$) and hence no evidence of non-linearity.

The type of non-linearity that a polynomial model is most sensitive to is constrained by the degree of the polynomial included in the model (i.e., a quadratic polynomial model will be most sensitive to departures from linearity that are roughly quadratic in nature). Therefore, further models involving higher polynomial terms could be examined. However, when the dataset includes multiple observed outcomes at the same level of the moderator (i.e., replicates), we can also conduct a more general lack-of-fit test that is much more flexible in terms of the types of non-linearity it can detect. For this, we need to extend the lack-of-fit test from regular regression (e.g., Fisher, 1922; Kutner et al., 2004) to the mixed-effects meta-regression setting, which can be done as follows.

We start with the linear model (i.e., $y_i = \beta_0 + \beta_1 x_{i1} + u_i + e_i$) and add dummy variables to the model for each level of the continuous moderator. However, this will lead to an overparameterized model. To obtain a model where all parameters are estimable, two of the dummy variables need to be removed, for example, those for the first and the last level (for the purposes of the test, it is completely arbitrary which two levels are removed). Therefore, if there are $j = 1, \ldots, \ell$ levels, then this yields the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \sum_{j=2}^{\ell-1} \alpha_j d_{ij} + u_i + e_i, \tag{11.22}$$

where $d_{ij} = 1$ if the ith outcome is at level j and 0 otherwise. This model can be fitted with the methods described in Section 11.2.5.

For the example data, we therefore include dummy variables corresponding to the 50, 75, 100, 125, 150, 175, and 200 mg levels of the dosage moderator (leaving out the 25 mg and 250 mg levels). The results for this model are shown in Table 11.3. The right-hand side

TABLE 11.3

Results for the Mixed-Effects Meta-Regression Model to Conduct the Lack-of-Fit Test

Term	Coefficient	Estimate	SE	z-value
Intercept	$\hat{\beta}_0$	−0.510	0.3035	−1.68
Dose (linear)	$\hat{\beta}_1$	0.008	0.0022	3.50
Dose (50 mg)	$\hat{\alpha}_1$	0.328	0.4916	0.67
Dose (75 mg)	$\hat{\alpha}_2$	−0.125	0.5257	−0.24
Dose (100 mg)	$\hat{\alpha}_3$	0.305	0.3433	0.89
Dose (125 mg)	$\hat{\alpha}_4$	0.328	0.3333	0.99
Dose (150 mg)	$\hat{\alpha}_5$	0.095	0.4135	0.23
Dose (175 mg)	$\hat{\alpha}_6$	0.354	0.5698	0.62
Dose (200 mg)	$\hat{\alpha}_7$	−0.221	0.4392	−0.50

of Figure 11.9 shows a scatterplot of the dosage moderator against the observed outcomes with the regression line from the linear model and the fitted values from the lack-of-fit model superimposed. As can be seen from the plot, (11.22) yields a "saturated" model that provides estimates of the average true log odds ratio for each level of the dosage moderator without any kind of implied shape. In fact, the lack-of-fit model is just a different parameterization of the model

$$y_i = \sum_{j=1}^{\ell} \alpha_j d_{ij} + u_i + e_i, \tag{11.23}$$

which includes a fixed effect for each level of the moderator variable. However, the advantage of the parameterization in (11.22) is that it allows for a direct test of the linearity assumption. In particular, using (11.8), we can conduct an omnibus test of $H_0\colon \alpha_2 = \cdots = \alpha_{\ell-1} = 0$. This yields $Q_M = 3.57$, which we compare against a chi-square distribution with seven degrees of freedom. This yields a p-value of 0.83 and hence no evidence that the relationship between dosage and outcomes is non-linear.

11.5.6 Checking the Normality Assumptions

Various assumptions underlying the models were discussed in Section 11.4, including the assumption that the sampling distributions are approximately normal. Let us assume that all of the assumptions discussed in that section are fulfilled. Still, this does not imply that the observed outcomes themselves are normally distributed. Even in the simplest case of the common-effect model, the y_i values are a mixture of normally distributed random variables with the same mean, θ, but different variances, s_i^2, which does not yield a normal marginal distribution. In addition, in models involving moderators, the marginal distribution is a mixture of variables with different means. Accordingly, there is no use in examining the distribution of the y_i values directly.

However, when a particular model indeed represents a rough but adequate approximation to the underlying data generating process, then this implies that the standardized

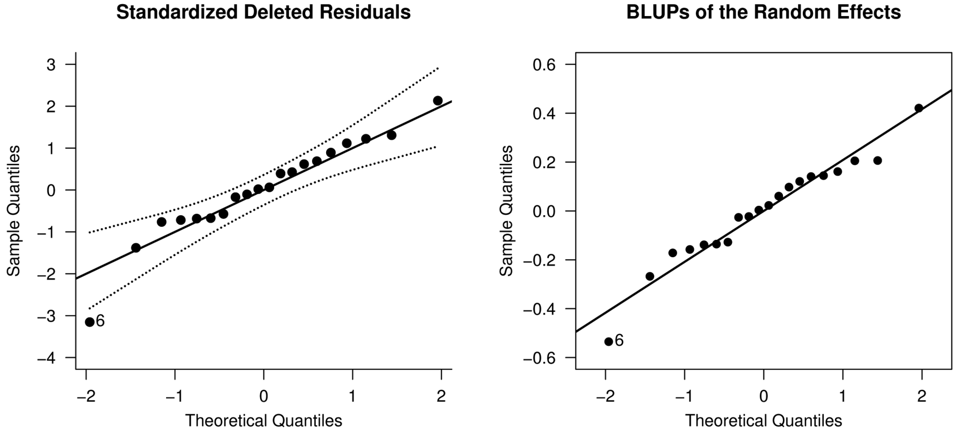


FIGURE 11.10
Q-Q plots of the standardized deleted residuals and the BLUPs of the random effects for the mixed-effects model.

(deleted) residuals should, at least approximately, follow a standard normal distribution. Q-Q plots can be used to examine whether this assumption holds. The left-hand side of Figure 11.10 shows such a plot of the standardized deleted residuals from the mixed-effects model. A diagonal reference line with an intercept of 0 and a slope of 1 was added to the plot. In addition, the dotted lines correspond to an approximate 95% pseudo confidence envelope, which was constructed based on the quantiles of sets of pseudo residuals simulated from the given model (for details, see Cook and Weisberg, 1982). Except for study 6, the points are roughly linear and fall close to the reference line. Therefore, there is no indication of non-normality in the standardized deleted residuals.

One other normality assumption not discussed so far underlies the random- and mixed-effects models. Besides assuming normally distributed sampling errors, the random- and mixed-effects models also make the additional assumption that the random effects, u_i, are normally distributed. In part, this assumption is often made purely because of convenience, that is, it greatly simplifies model fitting. However, this assumption can also be justified if we imagine that the (residual) heterogeneity in the true outcomes is a result of the influence of a large number of moderating factors, where each individual factor only has a small influence on the size of the true outcomes. When taken together, an approximately normal distribution could then emerge due to the central limit theorem. On the other hand, when outliers are present and/or when an important and strongly influential moderator has been omitted from the model, then this could lead to a non-normal random-effects distribution.

One possible approach to check this assumption is to compute the best linear unbiased predictions (BLUPs) of the random effects for a given model (Raudenbush and Bryk, 1985; Robinson, 1991) and then examine their distribution. Following the results in section (2.5), it can be shown that

$$\hat{u}_i = \lambda_i(y_i - \hat{\mu}_i) \tag{11.24}$$

provides predictions of the u_i values which have minimum mean squared error (among the class of linear unbiased estimators), where $\lambda_i = \hat{\tau}^2/(s_i^2 + \hat{\tau}^2)$. If the assumed model is correct, then the BLUPs should follow (approximately) a normal distribution.*

* There are two technical issues here. First, the computed values are really so-called empirical BLUPs (eBLUPs), since the unknown value of τ^2 is replaced by an estimate. Second, the eBLUPs do not have constant variance, so their marginal distribution may not be normal, even if the correct model is fitted and all assumptions hold.

The right-hand side of Figure 11.10 shows a Q-Q plot of the BLUPs for the mixed-effects model. The reference line again passes through the (0,0) point but has a slope equal to the observed standard deviation of the BLUPs (since the BLUPs do not have unit variance). Again, the outlying study 6 is quite noticeable, but otherwise, the points do not show any indication of a skewed or otherwise non-normal distribution.

However, diagnosing (non-)normality of the standardized (deleted) residuals and/or random effects in this manner is a difficult endeavor at best, especially when the number of studies included in the meta-analysis is small. Moreover, the distributions can be easily distorted when other assumptions are not fulfilled, when important moderators have been omitted from the model, or when the functional relationship between moderators and outcomes is misspecified. Finally, it is unclear how important it is to check the normality assumptions in the first place. For example, the assumption of normal errors is generally regarded as a relatively minor issue in the context of regular regression models (e.g., Gelman and Hill, 2006; Weisberg, 2006). Some simulations studies also indicate that meta-analytic models are quite robust to violations of the normality assumption of the random-effects distribution (Kontopantelis and Reeves, 2012; van den Noortgate and Onghena, 2003). However, further research is needed before more general recommendations can be made.

11.6 Checking for Influential Studies

So far, emphasis has been on assessing model fit and detecting outliers. Another issue to consider is the influence of each study on the results from the meta-analysis. Roughly speaking, an influential study is a study that exerts a considerable influence on the parameter estimates, test statistics, and/or conclusions that can be drawn from a given model. In some situations, certain findings (e.g., about the overall effectiveness of a treatment or the relevance of a particular moderator) may actually hinge on only one (or a few) of the studies in the dataset. In that case, it would be prudent to frame the corresponding conclusions more cautiously and to point out their volatility.

While it is often the case that outliers are also at least somewhat influential, it is important to properly distinguish between these concepts. For example, an outlier may not necessarily have much impact on the results if it comes from a very small study. Therefore, the presence of an outlier does not automatically call into question the conclusions drawn from the analyses. At the same time, a study whose observed outcome falls close to the fitted value based on the model (i.e., it is not an outlier) might still be influential, for example, if it is unusual in terms of its moderator values and its removal from the dataset might lead to considerable changes in any observed relationships.

An examination of a forest or scatterplot can already provide some indication whether influential studies may be present in a given dataset. However, as was the case for outliers, such informal approaches can be deceiving. We will therefore consider more rigorous methods for assessing and quantifying the influence of each study on various aspects of the fitted model.

11.6.1 Weights and Hat Values

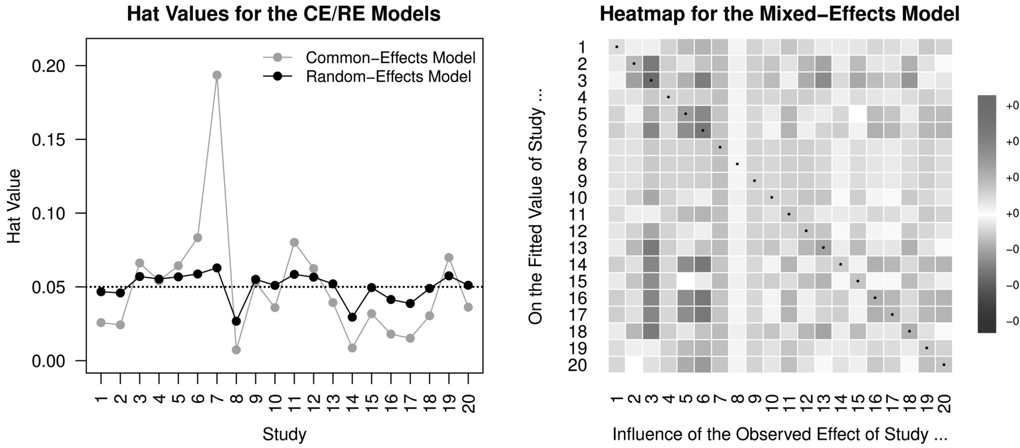
As described in Section 11.2.5, model fitting is done by means of weighted least squares, with weights equal to $w_i = 1/s_i^2$ for the common-effect and meta-regression models and $w_i = 1/(s_i^2 + \hat{\tau}^2)$ for the random- and mixed-effects models. An examination of the weights (either in their raw form or scaled to a percentage with $100\% \times w_i / \sum w_i$) can reveal which studies have the potential for exerting a strong influence on the results. However, in the context of a (mixed-effects) meta-regression model, the values of the moderator variables also play a prominent role in determining the potential influence of each study on the results. For example, when examining the relationship between the treatment dosage and the size of the outcomes in the example dataset, studies with very low or very high dosages will tend to be more influential than those with intermediate dosage levels.

Therefore, instead of just examining the weights, a more general approach is to compute the hat matrix $\boldsymbol{H}$ with (11.9). The values in the ith row of the hat matrix indicate how the fitted value of a particular study is a function of the observed values of all k studies (i.e., $\hat{\mu}_i = \sum_{j=1}^{k} h_{ij} y_j$, where h_{ij} is the jth value in the ith row of $\boldsymbol{H}$). Often, only the diagonal elements of $\boldsymbol{H}$ are examined (i.e., h_{ii} for $i, \ldots, k$, which is often abbreviated to just h_i; cf. Section 11.5.2), which are called the hat values (or "leverages") and which indicate to what extent the fitted value of a study is influenced by its own observed value. In fact, for the common- and random-effects models, each row of the hat matrix (and hence also the diagonal) is equal to $w_i / \sum w_i$, so for these models it is fully sufficient to examine the hat values and not the entire hat matrix. Moreover, as can be seen, the hat values for these models are simply the scaled weights.

However, for meta-regression models, the values in the hat matrix are not only a function of the weights, but also the moderator variables. Especially studies with unusual values for the moderator variable(s) (in essence, studies that are outliers in terms of the moderator space) will then tend to receive larger hat values. An interesting property of the hat values is that they always add up to $q+1$ (i.e., $\sum_{i=1}^{k} h_{ii} = q+1$). Hence, the reference value $(q+1)/k$ represents the (hypothetical) scenario where each study would have the same leverage on the results.

The left-hand side of Figure 11.11 shows the hat values for the common- and random-effects models for the example dataset, with the reference line drawn at $1/k$, corresponding to the case where each study has the same weight in the analysis. This is clearly not the case for the common-effect model. In particular, study 7 has a considerably larger leverage due to its relatively small sampling variance (and hence larger weight) compared with the rest of the studies. On the other hand, for the random-effects model, we see that the hat values are nearly equalized. This is a consequence of the relatively large estimate of τ^2 compared with the sampling variances of the studies, in which case the weights, and therefore the leverages, are dominated by $\hat{\tau}^2$. As a result, each individual study has approximately the same influence on the results from the random-effects model.

For the mixed-effects model, the right-hand side of Figure 11.11 provides a heatmap constructed on the basis of the entire hat matrix. The hat values are located along the diagonal and are marked with dots to make them easier to locate. Especially studies 3 and 6 have large leverages, which partly reflects their larger weights, but also their position in the moderator space (i.e., at the very low and high ends of the dosage continuum). Studies 14, 16, and 17 are equally extreme in terms of their dosages (all at the very low end), but their larger sampling variances (and hence lower weights) limits their leverages. By examining the corresponding rows of the hat matrix, we see that the fitted values for these studies are actually mostly a function of the observed outcomes of the other studies (especially



Hat Values for the CE/RE Models

Heatmap for the Mixed−Effects Model

FIGURE 11.11
Plot of the hat values for the common- and random-effects models and a heatmap of the hat matrix for the mixed-effects model.

studies 3, 5, and 6) and not their own. Finally, study 7, which had high leverage in the context of the common-effect model, now only plays a relatively minor role. Therefore, the potential impact of a study must be considered with respect to a specific model.

11.6.2 Cook's Distances

A high leverage study has the *potential* to exert considerable influence on the results but does not necessarily do so. A study actually becomes influential if the estimates obtained from the model change substantially depending on whether the study is included in the dataset or not. To determine whether this is the case for a particular study, we can make use of an approach introduced earlier where we examine the consequences of deleting the study from the dataset.

There are various aspects of a model that can be influenced by a study. Of particular interest are the model coefficients themselves (i.e., $\hat{\boldsymbol{\beta}}$ in (mixed-effects) meta-regression models or $\hat{\theta}$ and $\hat{\mu}$ in the common- and random-effects models). In order for a study with high leverage to become influential on this aspect of a model, its observed outcome must also deviate from the fitted value to a noteworthy degree, that is, it must be an outlier. As we have seen earlier, standardized deleted residuals are especially useful for detecting the latter. We can put these two ideas together and plot the leverages against the standardized deleted residuals. Influential studies will then be located at the top or bottom right-hand corner of the plot.

A measure that combines these two pieces of information into a single influence measure is Cook's distance (Cook and Weisberg, 1982; Viechtbauer and Cheung, 2010). It can be computed with

$$D_i = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}), \tag{11.25}$$

where $\hat{\boldsymbol{\beta}}_{(-i)}$ denotes the estimate of $\boldsymbol{\beta}$ computed with (11.6) when excluding the ith study from the model fitting. Written this way, the Cook's distance of a study can be interpreted

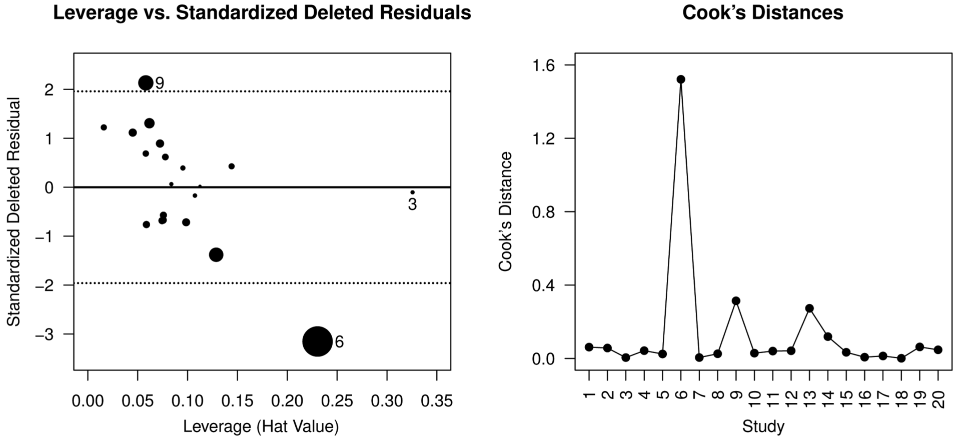


FIGURE 11.12
Scatterplot of the leverages versus the standardized deleted residuals and a plot of Cook's distances for the mixed-effects model.

as the Mahalanobis distance between the estimated model coefficients based on all k studies (i.e., $\hat{\boldsymbol{\beta}}$) and the estimates obtained when the study is excluded from the model fitting (i.e., $\hat{\boldsymbol{\beta}}_{(-i)}$). An equivalent way of expressing Cook's distance is

$$D_i = \sum_{j=1}^{k} \frac{(\hat{\mu}_j - \hat{\mu}_{j(-i)})^2}{s_j^2 + \hat{\tau}^2}, \tag{11.26}$$

which in turn can be interpreted as the Mahalanobis distance between the fitted values computed based on all k studies (i.e., $\hat{\mu}_j$ for $j=1,\ldots,k$) and the fitted values when the ith study is excluded from the model fitting (i.e., $\hat{\mu}_{j(-i)}$ for $j=1,\ldots,k$). Accordingly, a large value of D_i indicates notable changes in the model coefficients and fitted values depending on whether a study is included or removed from the dataset.

The left-hand side of Figure 11.12 shows a scatterplot of the leverages against the standardized deleted residuals for the mixed-effects model, with the points drawn proportional in size to Cook's distances. The right-hand side of the figure shows Cook's distances themselves. As we can see, high leverage points that are not outliers are not influential (study 3). Moreover, studies that are outliers but with relatively low leverage also do not exert much influence on the results (study 9). However, high leverage combined with poor fit results in an influential case (study 6).

11.6.3 Covariance Ratios and Other Deletion Diagnostics

Cook's distance quantifies the influence of each study on the model coefficients (and hence, the fitted values). However, this is not the only aspect of a model that can be affected by a study. In fact, a study may have relatively little influence on this part of a model, yet its removal can have other noteworthy consequences, which we would also want to be aware of.

For example, in random- and mixed-effects models, another important parameter besides the model coefficients is the estimate of (residual) heterogeneity. On the left-hand side of Figure 11.13, we can see a plot of the τ^2 estimates from the mixed-effects model

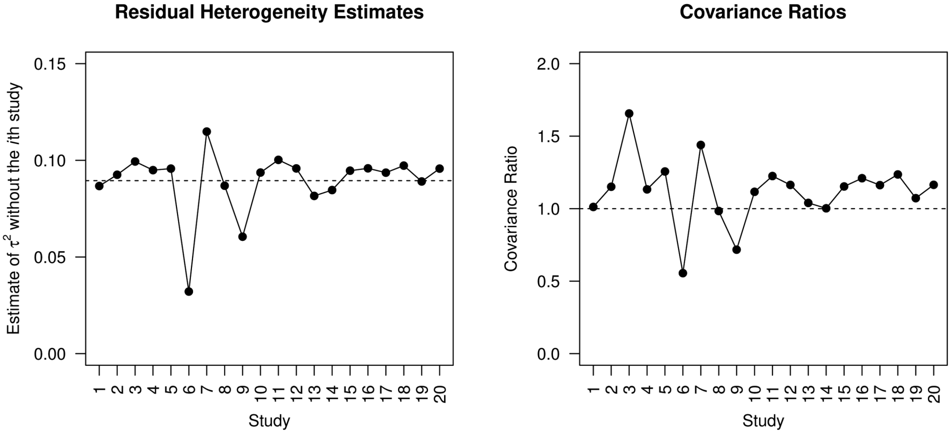


FIGURE 11.13
Plot of the leave-one-out estimates of τ^2 as each study is removed in turn and a plot of the covariance ratios for the mixed-effects model.

when each study is removed in turn. The horizontal dashed line corresponds to the estimate of τ^2 based on all k studies. Clearly, removal of study 6 leads to a substantial decrease in the amount of heterogeneity that is unaccounted for. This also applies to study 9, but to a lesser extent. On the other hand, removing study 7 would lead to a less pronounced but still discernible *increase* in the estimate.

Another aspect of all models worth considering is the precision with which we are able to estimate the model coefficients. At first sight, one would expect more data to lead to better (i.e., more precise) estimates, but this is not always the case. For random- and mixed-effects models, studies that introduce considerable (residual) heterogeneity into the data can actually lead to *decreases* in precision. At the same time, it is also informative to determine if there are studies in the dataset that are most responsible for driving up the precision of the estimates.

A useful measure for quantifying the effect of each study on this part of a model is the covariance ratio (Belsley et al., 1980; Viechtbauer and Cheung, 2010), given by

$$\text{COVRATIO}_i = \frac{\det\left[\text{Var}[\hat{\beta}_{(-i)}]\right]}{\det\left[\text{Var}[\hat{\beta}]\right]}, \tag{11.27}$$

where the numerator and denominator are generalized variances (i.e., the determinants of equation 7) for the reduced dataset (with the ith study excluded) and the full dataset, respectively. Since a smaller (generalized) variance is desirable, covariance ratios below one indicate that removal of a study leads to higher precision, while values above one indicate a decrease in precision.

The covariance ratios for the mixed-effects model are plotted on the right-hand side of Figure 11.13. Not surprisingly, removal of studies 6 and 9, which we identified earlier as sources of residual heterogeneity, would lead to increases in precision. Removing study 7, on the other hand, would have the opposite effect. Interestingly, study 3 also has a considerable covariance ratio. Recall that despite its high leverage, this study had essentially no influence on the model coefficients (i.e., its Cook's distance was very low). Moreover, it had no noteworthy effect on the estimate of τ^2. Yet, this study plays an important role, as its exclusion would result in substantially less precise estimates of the model coefficients.

Therefore, as this example demonstrates, some studies may only have a notable influence on this aspect of a model.

11.7 Other Types of Models

As discussed in Section 11.4, the models considered in this chapter assume that the sampling distributions of the observed outcomes are normal, that the observed outcomes are unbiased estimates, and that their sampling variances are known constants. Under certain circumstances, at least some of these assumptions are likely to break down.

An important case where we should be worried about violations of these assumption is in the context of meta-analyses examining the occurrence of rare events (e.g., Lane, 2013; see also Chapter 5). The sampling distributions of measures such as the risk difference and the log odds/risk ratio will then be poorly approximated by normal distributions. Moreover, estimates of the sampling variances will be very inaccurate, so that it is no longer acceptable to treat the variances as known constants. As a result of these assumption violations, inferential procedures (i.e., tests and confidence intervals) will no longer have nominal properties and the results/conclusions obtained cannot be trusted.

Fortunately, there is a wide variety of alternative models and methods available that can be used in this context, including Peto's method, the Mantel–Haenszel method, generalized linear mixed-effects models (i.e., mixed-effects logistic and Poisson regression), the non-central hypergeometric model, and the beta-binomial model (e.g., Mantel and Haenszel, 1959; Ma et al., 2016; Stijnen et al., 2010; Yusuf et al., 1985; see also Chapter 5). These methods relax certain underlying assumptions and try to model the observed data using more appropriate distributions. For example, for the data in Table 11.1, we could assume that x_i^T and x_i^C follow binomial distributions and then use logistic mixed-effects regression to model the log odds for remission in the treatment and control arms of each study. See Jackson et al., 2018 and Chapter 5 for more details.

Analogous model checking procedures as the ones described in this chapter can (and should) also be applied for such models. For example, akin to the tests for model misspecification described in Section 11.5.1, one can conduct likelihood ratio tests of $H_0\colon \tau^2 = 0$ in random- and mixed-effects logistic regression models. Outliers can again be detected by means of an examination of the residuals, although some additional complications arise in this context. For example, in logistic regression, we model the log odds in each study arm and hence the residuals will reflect deviations from the predicted log odds (or from the predicted event probabilities) for each arm. These residuals therefore do not directly address the question whether the log odds ratio of a particular study should be considered an outlier under a given model.

On the other hand, other methods generalize in a straightforward manner to logistic models. For example, GOSH plots could be generated based on the estimated model coefficients (although the computational burden would be increased considerably, especially when fitting random/mixed-effects logistic regression models), the lack of linearity test could be applied in the same manner, BLUPs of the random effects can be obtained and checked for normality, and influence measures such as Cook's distances and covariance ratios can be easily computed.

As discussed earlier, an assumption that applies specifically to random- and mixed-effects models concerns the nature of the random-effects distribution (i.e., the distribution of the underlying true effects). In practice, we typically assume that the random effects are normally distributed. Although it remains unclear how important it is to assess this assumption, it nevertheless can be worrisome if a diagnostic procedure (such as a Q-Q plot of the BLUPs) suggests considerable non-normality. In case non-normality is detected, one could explore models that allow for other types of distributional assumptions with respect to the random effects (Baker and Jackson, 2008; Beath, 2014; Lee and Thompson, 2008).

Finally, one additional assumption underlies all of the models described in Section 11.2 that warrants attention. As described, the models assume that the observed outcomes are independent. However, the independence assumption may be violated in various ways. For example, multiple outcomes extracted from the same sample of subjects are likely to be correlated. Returning to the example dataset, suppose that remission was defined and measured in two different ways in a particular study, leading to two log odds ratios that can be computed from the study. Then the sampling errors for the two outcomes are probably correlated (if patients who went into remission under the first definition also have higher chances of remission under the second definition). Similarly, if remission was measured at two different time points within a study, then the two corresponding log odds ratios are also probably correlated (if the occurrence of remission at the first time point is correlated with remission at the second time point).

Even if each study only provides a single outcome, it is still possible that the independence assumption is violated. For example, the underlying true outcomes may be correlated when multiple studies were conducted by the same author. Due to similarities in patient populations, methods, and procedures across studies (that are not captured by relevant moderator variables), the underlying true treatments effects may then be more alike for studies conducted by the same author than those conducted by different authors, which in essence again violates the independence assumption.

A common approach to deal with such dependencies is to reduce the dataset to one where dependencies are avoided (e.g., by selecting only one log odds ratio per study and/or author). Alternatively, multilevel and multivariate structures in a given dataset can be accounted for using appropriate models (e.g., Berkey et al., 1998; Jackson et al., 2011; Konstantopoulos, 2011). Model checking (including the detection of outliers and the identification of influential studies) also remains an important issue in the context of such analyses. Many of the methods discussed in this chapter can be generalized to such models, but the details of this are beyond the scope of this chapter.

11.8 Final Comments

Some final issues are worth commenting on. First of all, many meta-analyses involve only a relatively small number of studies. For example, a review of the Cochrane Database of Systematic Reviews indicated that the majority of Cochrane reviews contains only a handful of studies (Davey et al., 2011). That number tends to be somewhat higher for meta-analyses published in other outlets and/or for other disciplines (e.g., Cafri et al., 2010; Page et al., 2016), but meta-analyses with less than a dozen studies are still commonly encountered. Some of the techniques described in this chapter may be less informative or relevant in that context. At the same time, it is then even more important to check the data for outliers and influential studies, as their impact will tend to be larger in smaller datasets. Hence, standardized (deleted) residuals, Cook's distances, and other influence measures remain useful diagnostic tools. Still, the best one can hope to accomplish in such a situation is to identify cases where one particular study yields rather different results than the rest of the studies. If multiple studies yield very disparate results, then this will usually be indistinguishable from a situation where there is a high amount of heterogeneity in the data.

On the other hand, when *k* is large, it is important to realize that many of the techniques discussed in this chapter are "deletion diagnostics" that remove individual studies from the dataset and then examine the consequences thereof. These methods can also be effective for detecting multiple outliers and influential studies, but the presence of multiple outliers can lead to distortions such that none of the true outliers are detected (a phenomenon known as masking) or that some studies are incorrectly labeled as outliers (a phenomenon known as swamping) (Barnett and Lewis, 1978). Deletion diagnostics involving the simultaneous removal of multiple studies will then be required to sort out such intricacies. The GOSH plot may be especially useful in this context, as it is based on all possible subsets and hence does not require the specification of the number of studies to remove *a priori*.

For the most part, specific decision rules or cutoffs for deciding when a study should be considered an outlier and/or influential have been avoided throughout this chapter. Any such guidelines would ultimately be arbitrary (which also applies to the ±1.96 value to which the standardized deleted residuals were compared earlier). Instead, emphasis has been on visual inspection of the various diagnostic measures. By comparing the relative magnitude of the values across studies, one can often easily identify those studies that stand out with respect to a particular measure. The plot of Cook's distances (i.e., Figure 11.12) is exemplary for this approach.

An important point not discussed so far is the question what one should do when some studies are identified as potential outliers and/or influential. To begin with, one should check that the data are not contaminated in some way, for example, due to errors in reporting or coding. For example, a standard error of the mean that is mistaken for a standard deviation can lead to a substantial overestimate of the true standardized mean difference or the precision of a mean difference. In the context of the illustrative example, a study author may have reported a dosage of "50 mg tid" (for a daily dosage of 150 mg), but this may have accidentally been coded as a daily dosage of 50 mg.

However, in many cases, no such simple explanations will be found. In that case, one approach that one should never take is to simply remove the unusual studies from the analysis. All studies that fit the initial inclusion criteria should be reported and described. However, one may still probe the *robustness* of the analyses by excluding outliers or influential studies from the dataset. For example, is dosage still a significant when studies 3 and/or 6 are removed from the illustrative dataset? If the conclusion about the relevance of this moderator would be overturned, it would indicate that this finding is not particularly robust, and the finding should be described more cautiously. Fortunately, removal of either or both studies still yields the same conclusion in this example, which lends more credibility to the hypothesis that medication dosage is related to the treatment effectiveness.

Moreover, studies yielding unusual results may actually point toward (or rather, raise interesting new hypotheses about) conditions under which the effect or association of interest is particularly large or small (Hedges, 1986; Light and Pillemer, 1984). Such post-hoc explanations should, of course, be treated with caution, but they can be an opportunity to learn something new about the phenomenon being studied. Hence, it is actually in the best interest of the reviewer to apply some of the model checking methods described in the present chapter.

As a final practical point, it is worth noting that all of the methods described in this chapter are implemented in the R package `metafor` (Viechtbauer, 2010) (code corresponding to the analyses conducted is provided on the book website). General purpose software packages could also be used for model fitting and extracting diagnostic measures such as standardized (deleted) residuals, BLUPs, and Cook's distances (e.g., with PROC MIXED from SAS), but certain specialized plots for meta-analysis (e.g., forest, Baujat, and GOSH plots) are then not directly available.

References

Baker R and Jackson D, 2008. A new approach to outliers in meta-analysis. *Health Care Management Science* **11**(2): 121–131.

Barnett V and Lewis T, 1978. *Outliers in Statistical Data*. New York: Wiley.

Baujat B, Mahé C, Pignon J-P and Hill C, 2002. A graphical method for exploring heterogeneity in meta-analyses: Application to a meta-analysis of 65 trials. *Statistics in Medicine* **21**(18): 2641–2652.

Beath KJ, 2014. A finite mixture method for outlier detection and robustness in meta-analysis. *Research Synthesis Methods* **5**(4): 285–293.

Belsley DA, Kuh E and Welsch RE, 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F and Colditz GA, 1998. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* **17**(22): 2537–2550.

Berkey CS, Hoaglin DC, Mosteller F and Colditz GA, 1995. A random-effects regression model for meta-analysis. *Statistics in Medicine* **14**(4): 395–411.

Cafri G, Kromrey JD and Brannick MT, 2010. A meta-meta-analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research* **45**(2): 239–270.

Cook RD and Weisberg S, 1982. *Residuals and Influence in Regression*. London: Chapman and Hall.

Davey J, Turner RM, Clarke MJ and Higgins JP, 2011. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: A cross-sectional, descriptive analysis. *BMC Medical Research Methodology* **11**: 160.

DerSimonian R and Laird N, 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**(3): 177–188.

Field AP, 2001. Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods* **6**: 161–180.

Fisher RA, 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* **1**: 1–32.

Fisher RA, 1922. The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society* **85**(4): 597–612.

Friedrich JO, Adhikari NKJ and Beyene J, 2008. The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: A simulation study. *BMC Medical Research Methodology* **8**: 32.

Gelman A and Hill J, 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Hauck WW, Anderson S and Leahy FJ, 1982. Finite-sample properties of some old and some new estimators of a common odds ratio from multiple 2×2 tables. *Journal of the American Statistical Association* **77**(377): 145–152.

Hedges LV, 1981. Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics* **6**(2): 107–128.

LV Hedges. Estimation of effect size from a series of independent experiments. *Psychological Bulletin* **92**(2): 490–499, 1982a.

Hedges LV, 1982b. Fitting categorical models to effect size from a series of experiments. *Journal of Educational Statistics* **7**(2): 119–137.

Hedges LV, 1982c. Fitting continuous models to effect size data. *Journal of Educational Statistics* **7**(4): 245–270.

Hedges LV, 1986. Issues in meta-analysis. *Review of Research in Education* **13**: 353–398.

Hedges LV, 1992. Meta-analysis. *Journal of Educational Statistics* **17**(4): 279–296.

Higgins JPT and Thompson SG, 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, **21**(11): 1539–1558.

Hotelling H, 1953. New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society, Series B* **15**(2): 193–232.

Jackson D, Law M, Stijnen T, Viechtbauer W and White IR, 2018. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Statistics in Medicine* **37**(7): 1059–1085.

Jackson D, Riley R and White IR, 2011. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine* **30**(20): 2481–2498.

Konstantopoulos S, 2011. Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods* **2**(1): 61–76.

Kontopantelis E and Reeves D, 2012. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. *Statistical Methods in Medical Research* **21**(4): 409–426.

Kutner MH, Nachtsheim CJ, Neter J and Li W, 2004. *Applied Linear Statistical Models*. 5th Edition. New York: McGraw-Hill.

Lane PW, 2013. Meta-analysis of incidence of rare events. *Statistical Methods in Medical Research* **22**(2): 117–132.

Lee KJ and Thompson SG, 2008. Flexible parametric models for random-effects distributions. *Statistics in Medicine* **27**(3): 418–434.

Light RJ and Pillemer DB, 1984. *Summing Up: The Science of Reviewing Research*. Cambridge: Harvard University Press.

López-López JA, Marín-Martínez F, Sánchez-Meca J, van den NoortgateW and Viechtbauer W, 2014. Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. *British Journal of Mathematical and Statistical Psychology* **67**(1): 30–48.

Ma Y, Chu H and Mazumdar M, 2016. Meta-analysis of proportions of rare events: A comparison of exact likelihood methods with robust variance estimation. *Communications in Statistics, Simulation and Computation* **45**(8): 3036–3052.

Mantel N and Haenszel W, 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22(4): 719–748.

Olkin I, Dahabreh IJ and Trikalinos TA, 2012. Gosh – A graphical display of study heterogeneity. *Research Synthesis Methods* **3**(3): 214–223.

Olkin I and Pratt JW, 1958. Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics* **29**(1): 201–211.

Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco AC, Catalá-López F, Li L, Reid EK, Sarkis-Onofre R and Moher D, 2016. Epidemiology and reporting characteristics of systematic reviews of biomedical research: A cross-sectional study. *PLoS Medicine* **13**(5): e1002028.

Raudenbush SW, 2009. Analyzing effect sizes: Random-effects models. In Cooper H, Hedges LV and Valentine JC (Eds). *The Handbook of Research Synthesis and Meta-Analysis*. 2nd Edition. New York: Russell Sage Foundation, 295–315.

Raudenbush SW and Bryk AS, 1985. Empirical Bayes meta-analysis. *Journal of Educational Statistics* **10**(2): 75–98.

Robinson GK, 1991. That BLUP is a good thing: The estimation of random effects. *Statistical Science* **6**(1): 15–32.

Rücker G, Schwarzer G and Carpenter J, 2008. Arcsine test for publication bias in meta-analyses with binary outcomes. *Statistics in Medicine* **27**(5): 746–763.

Rücker G, Schwarzer G, Carpenter J and Olkin I, 2009. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine* **28**(5): 721–738.

Sánchez-Meca J and Marín-Martínez F, 2001. Meta-analysis of 2×2 tables: Estimating a common risk difference. *Educational and Psychological Measurement* **61**(2): 249–276.

Stijnen T, Hamza TH and Ozdemir P, 2010. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine* **29**(29): 3046–3067.

Thompson SG and Sharp SJ, 1999. Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine* **18**(20): 2693–2708.

van den Noortgate W and Onghena P, 2003. Multilevel meta-analysis: A comparison with traditional meta-analytic procedures. *Educational and Psychological Measurement* **63**(5): 765–790.

Viechtbauer W, 2010. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* **36**(3): 1–48.

Viechtbauer W and Cheung MW-L, 2010. Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods* **1**(2): 112–125.

Viechtbauer W, López-López JA, Sánchez-Meca J and Marín-Martínez F, 2015. A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods* **20**(3): 360–374.

Walter SD, 1985. Small sample estimation of log odds ratios from logistic regression and fourfold tables. *Statistics in Medicine* **4**(4): 437–444.

Walter SD and Cook RJ, 1991. A comparison of several point estimators of the odds ratio in a single 2×2 contingency table. *Biometrics* **47**(3): 795–811.

Weisberg S, 2006. *Applied Linear Regression*. 3rd Edition. Hoboken, NJ: Wiley.

Yusuf S, Peto R, Lewis J, Collins R and Sleight P, 1985. Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Disease* **27**(5): 335–371.