7

Weak instruments and finite-sample bias

In this chapter, we consider the effect of weak instruments on instrumental variable (IV) analyses. Weak instruments, which were introduced in Section 4.5.2, are those that do not explain a large proportion of the variation in the exposure, and so the statistical association between the IV and the exposure is not strong. This is of particular relevance in Mendelian randomization studies since the associations of genetic variants with exposures of interest are often weak. This chapter focuses on the impact of weak instruments on the bias and coverage of IV estimates.

7.1 Introduction

Although IV techniques can be used to give asymptotically unbiased estimates of causal effects in the presence of confounding, these estimates suffer from bias when evaluated in finite samples [Nelson and Startz, 1990]. A weak instrument (or a weak IV) is still a valid IV, in that it satisfies the IV assumptions, and estimates using the IV with an infinite sample size will be unbiased; but for any finite sample size, the average value of the IV estimator will be biased. This bias, known as weak instrument bias, is towards the observational confounded estimate. Its magnitude depends on the strength of association between the IV and the exposure, which is measured by the F statistic in the regression of the exposure on the IV [Bound et al., 1995]. In this chapter, we assume the context of 'one-sample' Mendelian randomization, in which evidence on the genetic variant, exposure, and outcome are taken on the same set of individuals, rather than subsample (Section 8.5.2) or two-sample (Section 9.8.2) Mendelian randomization, in which genetic associations with the exposure and outcome are estimated in different sets of individuals (overlapping sets in subsample, non-overlapping sets in two-sample Mendelian randomization).

We illustrate this chapter using data from the CRP CHD Genetics Collaboration (CCGC) to estimate the causal effect of blood concentrations of C-reactive protein (CRP) on plasma fibrinogen concentrations (Section 1.3). As the distribution of CRP is positively skewed, we take its logarithm and assume a linear relationship between log(CRP) and fibrinogen. Although log(CRP)

and fibrinogen are highly positively correlated ($r = 0.45$ to 0.55 in the examples below), it is thought that long-term elevated levels of CRP are not causally associated with an increase in fibrinogen.

We first demonstrate the direction and magnitude of weak instrument bias for IV estimates from real and simulated data (Section 7.2). We explain why this bias comes about, why it acts in the direction of the confounded observational association, and why it is related to instrument strength (Section 7.3). We discuss simulated results that quantify the size of this bias for different strengths of instruments and different analysis methods (Section 7.4). When multiple IVs are available, we show how the choice of IV affects the variance and bias of IV estimators (Section 7.5). We propose ways of designing and analysing Mendelian randomization studies to minimize bias (Section 7.6). We conclude with a discussion of this bias from both theoretical and practical viewpoints, ending with a summary of recommendations aimed at applied researchers on how to design and analyse a Mendelian randomization study to minimize bias from weak instruments (Section 7.7).

7.2 Demonstrating the bias of IV estimates

First, we demonstrate the existence and nature of weak instrument bias in IV estimation using both real and simulated data.

7.2.1 Bias of IV estimates in small studies

As a motivating example, we consider the Copenhagen General Population Study [Zacho et al., 2008], a cohort study from the CCGC with complete cross-sectional baseline data for 35 679 participants on CRP, fibrinogen, and three SNPs from the *CRP* gene region: rs1205, rs1130864, and rs3093077. We calculate the observational estimate by regressing fibrinogen on log(CRP), and the IV estimate by the two-stage least squares (2SLS) method using all three SNPs as IVs in a per allele additive model (Section 4.2.1). We then analyse the same data as if it came from multiple studies by dividing the data randomly into substudies of equal size, calculating estimates of association in each substudy, and combining the results using inverse-variance weighted fixed-effect meta-analysis. We divide the whole study into, in turn, 5, 10, 16, 40, 100, and 250 substudies. We recall that the F statistic from the regression of the exposure on the IV is used as a measure of instrument strength (Section 4.5.2).

We see from Table 7.1 that the observational estimate stays almost unchanged whether the data are analysed as one study or as several studies. However, as the number of substudies increases, the pooled IV estimate increases from near zero until it approaches the observational estimate. At the

Substudies	Observational estimate	2SLS IV estimate	Mean F statistic
1	1.68 (0.01)	-0.05 (0.15)	152.0
5	1.68 (0.01)	-0.01 (0.15)	31.4
10	1.68 (0.01)	0.09 (0.14)	16.4
16	1.68 (0.01)	0.23 (0.14)	10.8
40	1.68 (0.01)	0.46 (0.13)	4.8
100	1.67 (0.01)	0.83 (0.11)	2.5
250	1.67 (0.01)	1.27 (0.08)	1.6

TABLE 7.1
Estimates of effect (standard error) of log(CRP) on fibrinogen (μmol/l) from the Copenhagen General Population Study ($N = 35\,679$) divided randomly into substudies of equal size and combined using fixed-effect meta-analysis: observational estimates using unadjusted linear regression, IV estimates using 2SLS. Mean F statistics averaged across substudies from linear regression of log(CRP) on three genetic variants.

same time, the standard error of the pooled IV estimates decreases. We can see that even where the number of substudies is 16 and the average F statistic is around 10, there is a serious bias. The causal estimate with 16 substudies is positive ($p = 0.09$) despite the causal estimate with the data analysed as one study being near to zero.

7.2.2 Distribution of the ratio IV estimate

In order to investigate the distribution of IV estimates with weak instruments, we use a simulation exercise, taking a simple example of a confounded association with a single dichotomous IV [Burgess and Thompson, 2011]. Parameters are chosen such that the causal effect is null, but simply regressing the outcome on the exposure yields a strong positive confounded observational association of close to 0.5. We took 6 different values of the strength of the IV–exposure association, corresponding to mean F statistic values between 1.1 and 8.7.

Causal estimates are calculated using the ratio method, although with a single IV the estimates from the ratio, 2SLS and limited information maximum likelihood (LIML) methods are the same (Section 4.3.2). The resulting distributions for the estimate of the causal parameter are shown in Figure 7.1. For weaker IVs, there is a marked bias in the median of the distribution in the positive direction and the distribution of the IV estimate has long tails. For the weakest IV considered, the mean F statistic is barely above its null expectation of 1 and the median IV estimate is close to the confounded observational estimate of 0.5. For stronger IVs, the median of the distribution of IV estimates is close to zero. The distribution is skew with more extreme causal estimates tending to take negative values.

The analyses in Table 7.1 and simulations in Figure 7.1 show that IV estimates can be biased. This bias has two notable features: it is larger when the F statistic for the IV–exposure relationship is smaller, and it is in the direction of the confounded observational estimate.

7.3 Explaining the bias of IV estimates

We now try to provide some more intuitive understanding of why weak instrument bias occurs. We give three separate explanations for its existence, in terms of the definition of the ratio estimator, finite-sample violation of the IV assumptions, and sampling variation of IV estimators.

7.3.1 Correlation of IV associations

First, there is a correlation between the numerator (estimate of the G–Y association) and denominator (estimate of the G–X association) in the ratio estimator. To understand this, we consider a simple model of confounded association with causal effect β_1 of X on Y, with a dichotomous IV $G = 0$ or 1, and further correlation between X and Y due to association with a confounder U:

$$\begin{aligned} X &= \alpha_1 G + \alpha_2 U + \varepsilon_X \\ Y &= \beta_1 X + \beta_2 U + \varepsilon_Y \\ U &\sim \mathcal{N}(0, \sigma_U^2); \quad \varepsilon_X \sim \mathcal{N}(0, \sigma_X^2); \quad \varepsilon_Y \sim \mathcal{N}(0, \sigma_Y^2) \text{ independently.} \end{aligned} \tag{7.1}$$

We initially assume that $\sigma_X^2 = \sigma_Y^2 = 0$ for ease of explanation.

If $\bar{u}_j$ is the average confounder level for the subgroup with $G = j$ (where $j = 0, 1$), an expression for the causal effect from the ratio method is:

$$\beta_1^R = \frac{\Delta Y}{\Delta X} = \frac{\beta_1 \Delta X + \beta_2 \Delta U}{\Delta X} = \beta_1 + \frac{\beta_2 \Delta U}{\alpha_1 + \alpha_2 \Delta U} \tag{7.2}$$

where $\Delta U = \bar{u}_1 - \bar{u}_0$ is normally distributed with expectation zero; ΔX and ΔY are defined similarly. When the instrument is strong, α_1 is large compared to $\alpha_2 \Delta U$. Then the expression β_1^R will be close to β_1. When the instrument is weak, α_1 may be small compared to $\beta_2 \Delta U$ and $\alpha_2 \Delta U$. Then the bias $\beta_1^R - \beta_1$ is close to $\frac{\beta_2}{\alpha_2}$, which is approximately the bias of the confounded observational association (it is exactly this if α_1 is zero). This is true whether ΔU is positive or negative. Figure 7.2 (top panel) shows how the IV estimate bias varies with ΔU. Although for any non-zero α_1 the IV estimator will be an asymptotically consistent estimator as sample size increases and ΔU tends towards zero, a bias in the direction of the confounded association will be present in finite samples. From Figure 7.2 (top panel), the median bias will be positive, as the

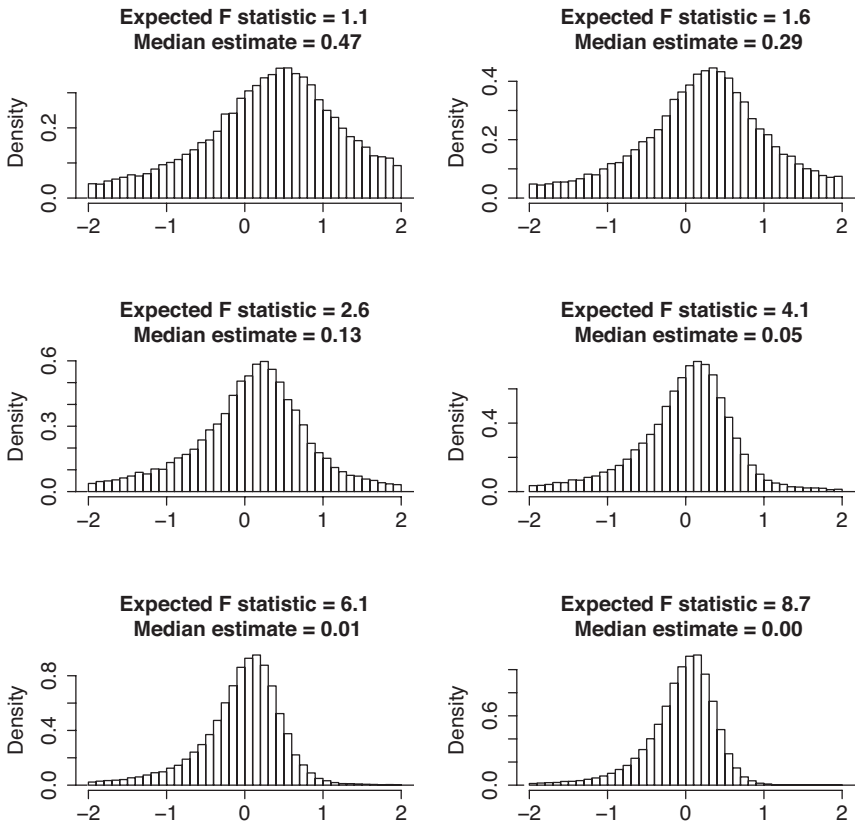

FIGURE 7.1
Histograms of IV estimates of a null causal effect using weak instruments from simulated data for six strengths of the IV–exposure association. Average F statistics and median IV estimates for each scenario are shown.

estimate is greater than β_1 when $\Delta U > 0$ or $\Delta U < -\frac{\alpha_1}{\alpha_2}$, which happens with probability greater than 0.5.

This also explains the heavier negative tail in the histograms in Figure 7.1. The estimator takes extreme values when the denominator $\alpha_1 + \alpha_2\,\Delta U$ is close to zero. Taking parameters α_1, α_2 and β_2 as positive, as in the example of Section 7.2.2, this is associated with a negative value of ΔU, whence the numerator $\beta_2\,\Delta U$ will be negative. As ΔU has expectation zero, the denominator is more likely to be small and positive than small and negative, giving more negative extreme values of β_1^R than positive ones.

If there is independent error in X and Y (that is, σ_X^2 and σ_Y^2 in equation (7.1) are non-zero), then the picture is similar, but more noisy, as seen in Figure 7.2 (bottom panel). The expression for the IV estimator is:

$$\beta_1^R = \beta_1 + \frac{\beta_2\,\Delta U + \Delta\varepsilon_Y}{\alpha_1 + \alpha_2\,\Delta U + \Delta\varepsilon_X}$$

where $\Delta\varepsilon_X = \bar{\varepsilon}_{X1} - \bar{\varepsilon}_{X0}$ and $\Delta\varepsilon_Y = \bar{\varepsilon}_{Y1} - \bar{\varepsilon}_{Y0}$ are defined analogously to ΔU above.

7.3.2 Finite-sample violation of IV assumptions

An alternative explanation of weak instrument bias is in terms of violation of the second IV assumption in a finite sample. Although a valid instrument will be asymptotically independent of all confounders, in a finite sample there will be a non-zero correlation between the instrument and confounders. This correlation biases the IV estimator towards the observational confounded association.

If the instrument is strong, then the difference in mean exposure between genetic subgroups will be mainly due to the genetic instrument, and the difference in outcome (if any) will be due to this difference in exposure. However if the instrument is weak, that is it explains little variation in the exposure, the chance difference in confounders may explain more of the difference in mean exposure between genetic subgroups than the instrument. If the effect of the instrument is near zero, then the estimate of the "causal effect" approaches the association between exposure and outcome resulting from changes in the confounders, which is the observational confounded association [Bound et al., 1995].

7.3.3 Sampling variation within genetic subgroups

Finally, we offer a graphical explanation of weak instrument bias. To do this, we simulate data with a negative causal effect of the exposure on the outcome, but with positive confounding giving a strong positive observational association between the exposure and outcome. We generate 1000 simulated datasets with 600 subjects divided equally into three genetic subgroups

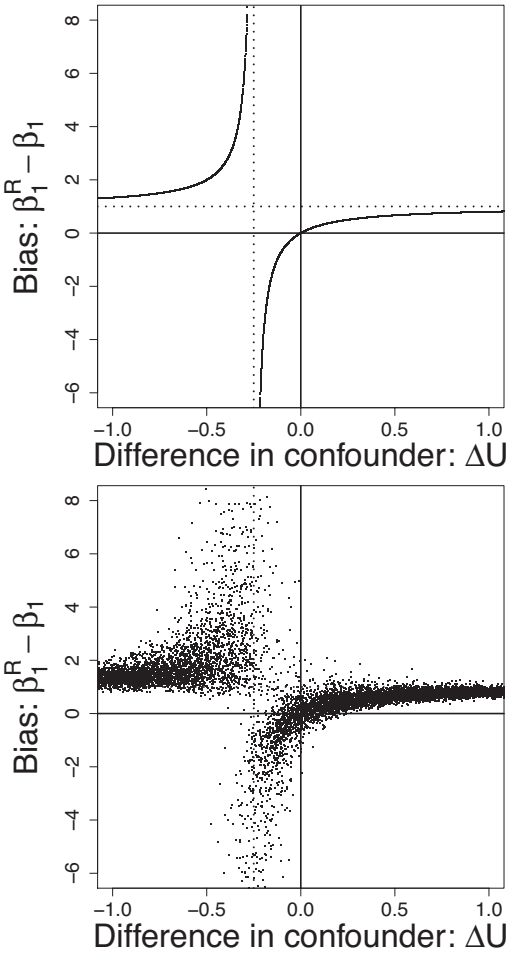

FIGURE 7.2
Bias in IV estimator as a function of the difference in mean confounder between groups ($\alpha_1 = 0.25$, $\alpha_2 = \beta_2 = 1$). Horizontal dotted line is at the confounded association $\frac{\beta_2}{\alpha_2}$, and the vertical dotted line at $\Delta U = -\frac{\alpha_1}{\alpha_2}$ where β_1^R is not defined. Top panel: no independent error in X or Y; bottom panel: $\Delta\varepsilon_X, \Delta\varepsilon_Y \sim \mathcal{N}(0, 0.1^2)$ independently.

($G = 0$, 1, or 2):

$$\begin{aligned} x_i &= \alpha_1 g_i + u_i + \varepsilon_{Xi} \\ y_i &= \beta_1 x_i + u_i + \varepsilon_{Yi} \\ u_i &\sim \mathcal{N}(0, \sigma_U^2); \varepsilon_{Xi} \sim \mathcal{N}(0, \sigma_X^2); \varepsilon_{Yi} \sim \mathcal{N}(0, \sigma_Y^2) \text{ independently.} \end{aligned} \tag{7.3}$$

We set $\beta_1 = -0.4$, $\sigma_U^2 = 1^2$, $\sigma_X^2 = 0.2^2$, and $\sigma_Y^2 = 0.2^2$, and take four values for the strength of the IV ($\alpha_1 = 0.5$, 0.2, 0.1, and 0.05) corresponding to expected F statistics of 100, 16, 4.7, and 2.0. The mean levels of exposure and outcome for each genetic subgroup from each simulated dataset are plotted (Figure 7.3), representing joint density functions for each subgroup. To examine the sampling distribution of the IV estimate, we draw one point at random from each of these distributions; the gradient of the line through these three points is the 2SLS IV estimate. When the instrument is strong, the large differences in exposure between the subgroups due to variation in the IV will generally lead to estimating a negative effect of exposure on outcome. When the instrument is weak, the differences in exposure between the subgroups due to the IV are small and the positively confounded observational association is more likely to be recovered.

7.4 Properties of IV estimates with weak instruments

In the previous section, we showed that IV estimates are biased in finite samples. In this section, we consider the magnitude of the bias in IV estimates, as well as the coverage of IV methods with weak instruments.

7.4.1 Bias of IV estimates

The bias of an estimator is the difference between the expectation of the estimator and the true value of the parameter. In IV analysis, the relative mean bias is the ratio of the bias of the IV estimator ($\hat{\beta}_{IV}$) to the bias of the observational association ($\hat{\beta}_{OBS}$) found by linear regression of the outcome on the exposure:

$$\text{Relative mean bias} = \frac{\mathbb{E}(\hat{\beta}_{IV}) - \beta_1}{\mathbb{E}(\hat{\beta}_{OBS}) - \beta_1}. \tag{7.4}$$

The relative mean bias from the 2SLS method is asymptotically approximately equal to $1/\mathbb{E}(F)$, where $\mathbb{E}(F)$ is the expected F statistic in the regression of the exposure on the IV [Staiger and Stock, 1997]. This approximation is only valid when the number of IVs is at least three. The rule-of-thumb of $F < 10$ indicating weak instruments (Section 4.5.2) derives from this expression. This rule approximately limits the bias in the IV estimate to less than

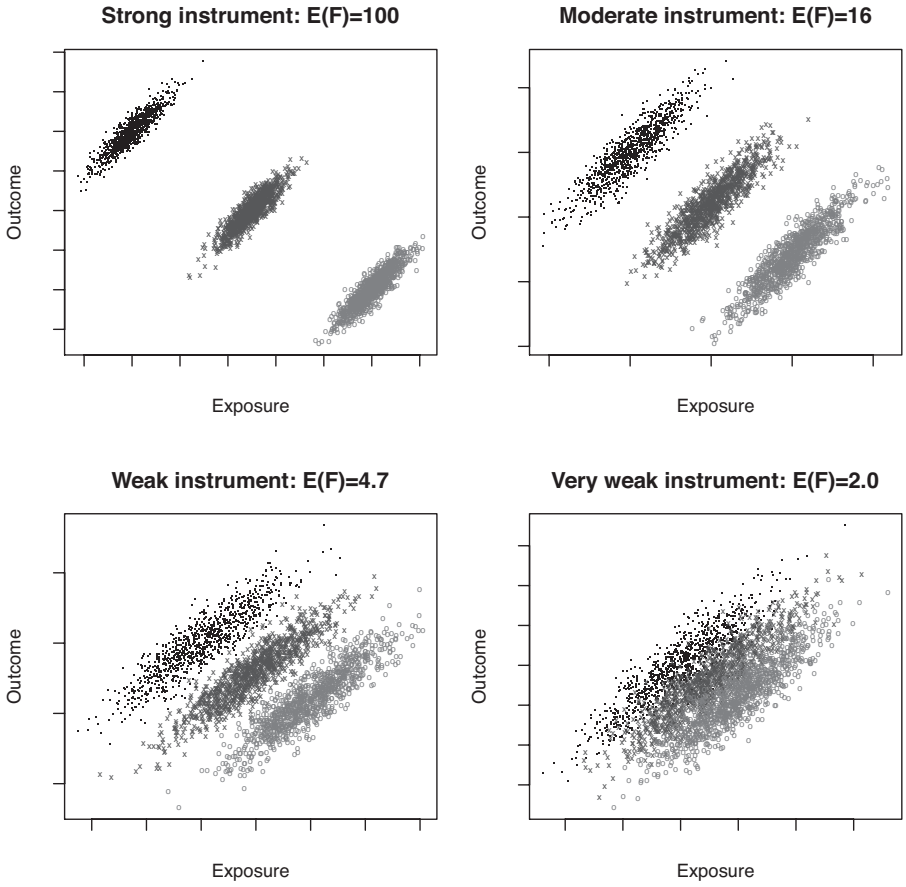

FIGURE 7.3
Distribution of mean outcome and mean exposure levels in three genetic subgroups (indicated by different symbols and shades of grey) for various strengths of the instrument, with expected values of the F statistic. One point of each colour comes from each of 1000 simulated datasets. The IV estimate in each simulation is the gradient of the line through the three points.

10% of the bias in the observational association estimate. However, weak instrument bias depends in a graded way on the F statistic, and such cut-offs are not always helpful or sensible. Biases less than this, corresponding to greater F statistics, can be important in practice. Moreover, as explained later in this chapter, there is an important distinction between the expected F statistic, on which the magnitude of the bias depends, and the F statistic observed in a particular dataset.

With a single IV, the expected value of the 2SLS estimate, and hence the bias, is undefined (Section 4.1.6). Simulations have shown that the median bias of the 2SLS method with a single IV (or equivalently the ratio or LIML method) is close to zero even for IVs with expected F statistics around 5, where the median bias is defined as the difference between the median estimate and the true value [Burgess and Thompson, 2011].

Other methods, such as likelihood-based methods, are less susceptible to bias. Although the mean bias of the LIML estimate is undefined (Section 4.3.2), the median bias is close to zero [Angrist and Pischke, 2009]. Simulations for Bayesian methods for IVs with expected F statistics around 5 have shown mean and median bias close to zero [Burgess and Thompson, 2012].

7.4.2 Coverage of IV estimates

In addition to problems of bias, IV estimates with weak instruments can have underestimated coverage [Stock and Yogo, 2002; Mikusheva and Poi, 2006]. As seen in Figure 7.1, the distribution of the IV estimate has long tails, and so is poorly approximated by a normal distribution. This means that asymptotically derived confidence intervals may underestimate the true uncertainty in the causal effect. This underestimation is especially severe when confounding is strong. Simulations for the 2SLS method have shown coverage as low as 75% for a nominal 95% confidence interval [Burgess and Thompson, 2012]. Similar results have been observed for the LIML method when there is a large number of IVs; while a correction is available (Bekker standard errors [Bekker, 1994]), this leads to inefficient estimates [Davies et al., 2014]. Confidence intervals from Fieller's theorem (Section 4.1.5), which are not constrained to be symmetric (or even finite), or those which do not rely on asymptotic assumptions, such as credible intervals from a Bayesian posterior distribution drawn from Monte Carlo Markov chain (MCMC) sampling, result in better coverage properties [Imbens and Rosenbaum, 2005]. Alternatively, confidence intervals from inverting a test statistic, such as the Anderson–Rubin test statistic [Anderson and Rubin, 1949] or the conditional likelihood ratio test statistic [Moreira, 2003] give appropriate confidence levels under the null hypothesis with weak instruments [Mikusheva, 2010].

7.4.3 Lack of identification

For semi-parametric approaches to IV analysis, such as the generalized method of moments (GMM) or structural mean models (SMM), there is no guarantee that a unique parameter estimate will be obtained, as the estimating equations may have no or multiple solutions (Section 4.4.3*). This is a common problem when the instrument is weak. Even when there is a unique solution, if the gradient of the graph of the objective function from the estimating equations against parameter values is close to zero in the neighbourhood of the parameter estimate, or if the objective function cannot be well approximated by a quadratic function, then identification is said to be weak, and problems of bias and coverage as explained above are likely to occur. Simulations suggest that the probability of obtaining a unique solution to the estimating equations with a binary outcome and a log-linear model is not especially sensitive to the sample size, depending more on the coefficient of determination (R^2, the proportion of variance in the exposure explained by the IV(s)) than the F statistic. With R^2 of 2% or less, lack of identification in a multiplicative GMM (or equivalently a multiplicative SMM) model was observed in over 50% of simulated datasets even when the F statistic was in the hundreds or even thousands [Burgess et al., 2014c].

7.5 Bias of IV estimates with different choices of IV

Including more instruments, where each instrument explains extra variation in the exposure, should give more information on the causal parameter (see Chapter 8). However, bias may increase, due to the weakening of the set of instruments. In this section, we consider the impact of choice of instrument on the bias of IV estimates.

7.5.1 Multiple candidate IVs in simulated data

In order to investigate how using multiple instruments affects the bias of IV estimates, we perform simulations in a model [Burgess and Thompson, 2011] where, for each participant indexed by i, the exposure x_i depends linearly on six dichotomous IVs ($g_{ik}, k = 1, \ldots, 6$), a normally distributed confounder u_i, and an independent normally distributed error term ε_{Xi}. Outcome y_i is a linear combination of exposure, confounder, and an independent error term

ε_{Yi}:

$$x_i = \sum_{k=1}^{6} \alpha_{1k}\, g_{ik} + \alpha_2\, u_i + \varepsilon_{Xi} \tag{7.5}$$
$$y_i = \beta_1\, x_i + \beta_2\, u_i + \varepsilon_{Yi}$$
$$u_i, \varepsilon_{Xi}, \varepsilon_{Yi} \sim \mathcal{N}(0, 1^2) \text{ independently.}$$

We set $\beta_1 = 0, \alpha_2 = 1, \beta_2 = 1$ so that X is observationally strongly positively associated with Y, but the causal effect is null. We take parameters for the genetic association $\alpha_{1k} = 0.4$ for each genetic instrument k, corresponding to a mean F statistic of 10.2. We used a sample size of 512 divided equally between the $2^6 = 64$ genetic subgroups. The IVs are uncorrelated, so that the variation in X explained by each IV is independent, and the mean F statistics do not depend greatly on the number of IVs (mean 10.2 using 1 IV, 11.3 using 6 IVs).

Table 7.2 shows the median and 95% range of the estimates from the 2SLS and LIML methods and the mean estimate for the 2SLS method using all combinations of all numbers of IVs as the instrument, with the mean across simulations of the F statistic for all the instruments used. We also give results using the IV with the greatest and lowest observed F statistics in each simulation, as well as using all IVs with an F statistic greater than 10 in univariate regressions of exposure on each IV.

Using 2SLS, as the number of IVs increases, the bias increases, despite the mean F statistic remaining fairly constant. This is because there is a greater risk of imbalances in confounders between the greater number of genetic subgroups defined by the instruments. The data are being subdivided in more different ways, and so there is more chance of these divisions giving genetic subgroups with different average levels of confounders. However, the variability of the IV estimator decreases. This is because a greater proportion of the variance in the exposure is modelled. The greatest increase in median bias is from one IV to two IVs, and coincides with the greatest increase in precision. With the 2SLS method, we therefore have a bias–variance trade-off in deciding how many IVs to use [Zohoori and Savitz, 1997].

While LIML provides estimates which are slightly more variable than 2SLS, a similar increase in precision with the number of IVs is observed, but no increase in bias. For 2SLS, the mean estimates are slightly smaller than the median estimates presented. In the case of a single IV, the theoretical mean is infinite (Section 4.1.6). For LIML, the mean bias is infinite for all numbers of IVs (Section 4.3.2).

Using the single IV with the greatest F statistic gives markedly biased results, despite a mean F statistic of 23.9. There is a similar bias only using IVs with $F > 10$. In the simulation, each IV in truth explains the same amount of variation in the exposure. If the IVs are chosen to be included in an analysis because they explain a large proportion of the variation in the exposure in the data under analysis, then the estimate using these IVs is

additionally biased. This is because the IVs explaining the most variation will be overestimating the proportion of true variation explained, due to chance correlation with confounders. In the notation of Section 7.3.1, ΔU is large and, having the same sign as α_1, leads to an estimate biased in the direction of $\frac{\beta_2}{\alpha_2}$. Conversely, if the IV with the least F statistic is used as an instrument, the IV estimator will be biased in the opposite direction to the observational association, as shown in Table 7.2.

So we see that if the F statistic is used either to choose between instruments, or via a rule such as only including an IV in the analysis if $F > 10$, this procedure itself introduces a selection bias which can be greater in magnitude than the bias from weak instruments [Hall et al., 1996]. In a more realistic example, IVs would not all have the same true strength. However, the large sampling variation in F statistics means that choosing between IVs on the basis of a single measured F statistic is unreliable. One solution to this in practice is to use the strength of the IVs in an independent dataset to determine the IVs to include in an applied analysis, or to use an allele score to summarize multiple variants as a single IV (see Chapter 8).

	Median	2.5% to 97.5% quantiles			Mean	Mean F
IVs used	2SLS		LIML		2SLS [1]	statistic
1 IV	0.00	−1.12 to 0.53			–	10.2
2 IVs	0.02	−0.54 to 0.39	0.00	−0.64 to 0.39	0.00	10.4
3 IVs	0.03	−0.39 to 0.33	0.00	−0.48 to 0.32	0.02	10.6
4 IVs	0.03	−0.31 to 0.30	0.00	−0.40 to 0.28	0.02	10.8
5 IVs	0.04	−0.26 to 0.27	0.00	−0.34 to 0.26	0.03	11.0
6 IVs	0.04	−0.23 to 0.26	0.00	−0.31 to 0.23	0.03	11.3
Greatest F	0.14	−0.30 to 0.52			–	23.9
Least F	−0.32	−2.57 to 0.58			–	6.7
IVs with $F > 10$	0.11	−0.20 to 0.39	0.10	−0.22 to 0.39	0.11	16.4

TABLE 7.2
Evaluation of bias: Median and 95% range of estimates of $\beta_1 = 0$ using 2SLS and LIML methods, mean estimate using 2SLS method and mean F statistic across 100 000 simulations using combinations of six uncorrelated instruments, using the instrument with the greatest/least F statistic, and using all instruments with univariate F statistics greater than 10.

[1] Mean estimate is reported only when it is not theoretically infinite

7.5.2 Multiple candidate IVs in the Framingham Heart Study

As a further illustration, we consider the Framingham Heart Study, a cohort study measuring CRP and fibrinogen at baseline with complete data on 1500 participants for nine SNPs in the *CRP* gene. The observational estimate of the log(CRP)–fibrinogen (μmol/l) association is 1.13 (95% CI 1.05 to 1.22). We calculate the causal estimate of the association using the 2SLS method with different numbers of SNPs as an instrument, using a per allele additive model. Figure 7.4 shows a plot of the 2SLS IV estimates against number of instruments, where each point represents the causal estimate calculated using the 2SLS method with a different combination of SNPs. The range of point estimates of the causal effect reduces as we include more instruments, but the median causal estimate across the different combinations of IVs increases. The 2SLS estimate using all nine SNPs in an per allele additive model is -0.01 (95% CI: -0.72 to 0.71, $p = 0.99$, $F_{9,1490} = 3.34$). If we relax the assumptions of a per allele genetic model with additivity between SNPs to instead use a fully saturated model with one coefficient for each of the 49 genotypes represented in the data, the 2SLS estimate is 0.79 (95% CI 0.42 to 1.16, $p < 0.001$, $F_{48,1451} = 1.66$). Using LIML, the estimate from the saturated genetic model is 0.05 (95% CI -0.71 to 0.81, $p = 0.89$) – much less biased than the 2SLS estimate.

This illustrates the bias in the 2SLS method due to the use of multiple instruments, showing how an estimate close to the observational association can be obtained by injudicious choice of instrument. In the extreme case, if each of the individuals in a study were placed into separate genetic subgroups, then the IV estimate would be exactly the observational association. The LIML method with the saturated genetic model gives a substantially different answer to the 2SLS method, an indication that the 2SLS estimate may be biased.

7.6 Minimizing the bias of IV estimates

To provide guidance for epidemiological applications, we now list specific ways by which bias from weak instruments can be minimized in the design and analysis of Mendelian randomization studies.

7.6.1 Increasing the F statistic

As stated previously, the bias in 2SLS IV estimates depends on the expected F statistic in the regression of the exposure on the IV. This means that bias can be reduced by increasing the expected F statistic. The F statistic is

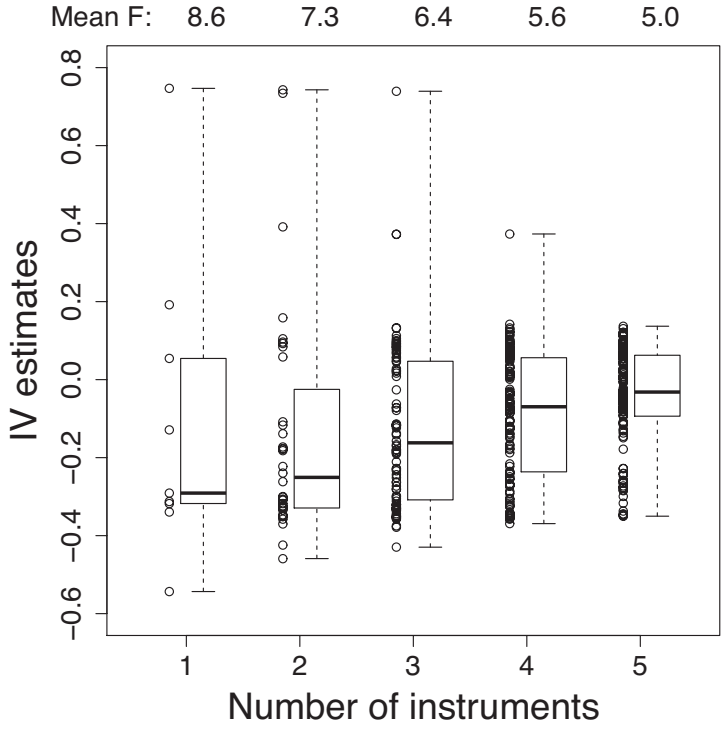

FIGURE 7.4
2SLS IV estimates for causal effect in the Framingham Heart Study of log(CRP) on fibrinogen (μmol/l) using all combinations of varying numbers of SNPs as IVs. Point estimates, associated box plots (median, inter-quartile range, range) and mean F statistics across combinations are displayed.

related to the proportion of variance in the exposure explained by the genetic variants (R^2), sample size (N) and number of instruments (K) by the formula $F = (\frac{N-K-1}{K})(\frac{R^2}{1-R^2})$. As the F statistic depends on the sample size, bias can be reduced by increasing the sample size. Similarly, if there are instruments that are not contributing much to explaining the variation in the exposure, then excluding these instruments will increase the F statistic. In general, employing fewer degrees of freedom to model the genetic association, that is using parsimonious models, will increase the F statistic and reduce weak instrument bias, provided that the model does not misrepresent the data [Pierce et al., 2011; Palmer et al., 2011a]. Simulations have shown that, even when the true model is only approximately linear in the IV, a per allele genetic model reduces bias [Burgess and Thompson, 2011].

However, it is not enough to simply rely on an F statistic measured from data to inform us about bias [Hall et al., 1996]. Returning to the example from Section 7.2.1 where we divided the Copenhagen General Population Study into 16 equally sized substudies with mean F statistic 10.8, Figure 7.5 shows the estimates of these 16 substudies using the 2SLS method with their corresponding F statistics. We see that the substudies which have greater estimates are the ones with larger F statistics; the correlation between F statistics and point estimates is 0.83. The substudies with higher F statistics also have tighter CIs and so receive more weight in the meta-analysis. If we exclude from the meta-analysis substudies with an F statistic less than 10, then the pooled estimate increases from 0.23 (SE 0.14, $p = 0.09$) to 0.43 (SE 0.16, $p = 0.006$). Equally, if we only use as instruments in each substudy the IVs with an F statistic greater than 10 when regressed in a univariate regression on the exposure, then the pooled estimate increases to 0.28 (SE 0.15, $p = 0.06$). So neither of these approaches are useful in reducing bias.

Although the expectation of the F statistic is a good indicator of bias, the observed F statistic shows considerable variation. In the 16 substudies of Figure 7.5, the measured F statistic ranges from 3.4 to 22.6. In more realistic examples, assuming similar instruments in each study, larger studies would have higher expected F statistics which would correspond to truly stronger instruments and less bias. However, the sampling variation of causal effects and observed F statistics in each study would still tend to follow the pattern of Figure 7.5, with larger observed F statistics corresponding to more biased causal estimates.

So while it is desirable to use strong instruments, the measured strength of instruments in data is not a good guide to the true instrument strength. Echoing the comments of Section 7.5 regarding the inclusion of IVs in a model, any guidance that relies on providing a threshold (such as $F > 10$) as an inclusion criterion is flawed and may introduce more bias than it prevents.

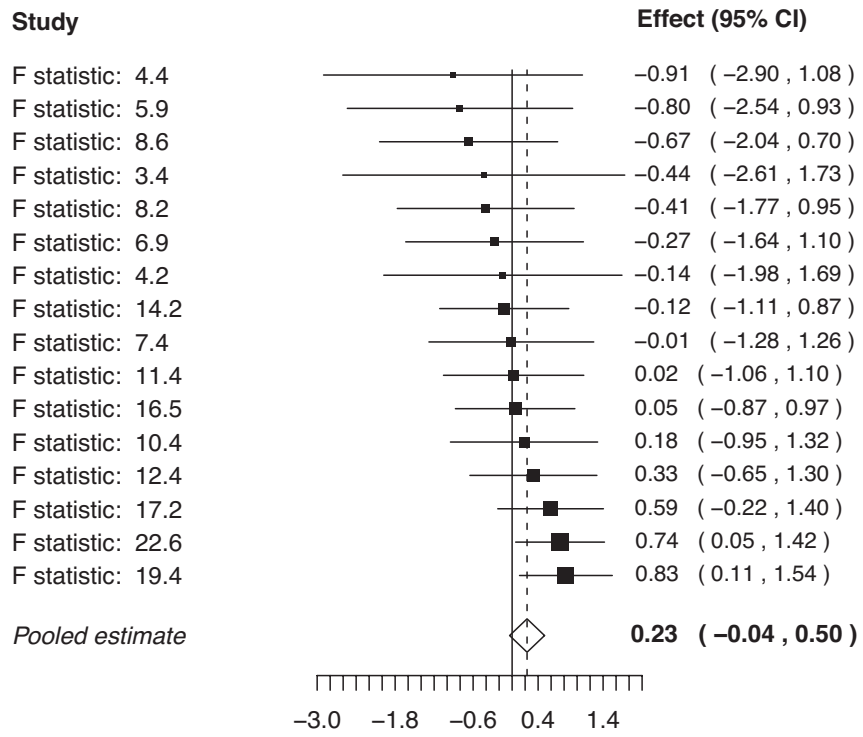

FIGURE 7.5
Forest plot of causal estimates of log(CRP) on fibrinogen (μmol/l) using data from the Copenhagen General Population Study divided randomly into 16 equally sized substudies (each $N \simeq 2230$). Studies ordered by causal estimate. F statistic from regression of exposure on three IVs. Size of markers is proportional to weight in a fixed-effect meta-analysis.

7.6.2 Adjustment for measured covariates

If we can find measured covariates that explain variation in the exposure, and that are not on the causal pathway between exposure and outcome, then we can incorporate these covariates in our model. This will increase precision in the genetic association with the exposure and reduce weak instrument bias. Simulations have shown that we may also see an increase in the precision of the IV estimator if these covariates are additionally used to explain variation in the outcome [Burgess et al., 2011b].

As an example, we consider data on interleukin-6 (IL6), a cytokine which is involved in the inflammation process upstream of CRP and fibrinogen [Hansson, 2005]. Elevated levels of IL6 lead to elevated levels of both CRP and fibrinogen, so IL6 is correlated with short-term variation in CRP [Kaptoge et al., 2010], but is independent of underlying genetic variation in CRP [CCGC, 2011]. We assume that it is a confounder in the association of CRP with fibrinogen and not on the causal pathway (if such a pathway exists). As IL6 has a positively skewed distribution, we take its logarithm.

We use data from the Cardiovascular Health Study, a cohort study from the CCGC measuring CRP, IL6 and fibrinogen at baseline, as well as three SNPs (rs1205, rs1417938, and rs1800947) on the CRP gene, with complete data for 4137 subjects. The proportion of variance in log(CRP) explained by log(IL6) is 26%. We calculate the 2SLS IV estimate of the CRP–fibrinogen association for each SNP separately and for all the SNPs together in an per allele additive model, both without and with adjustment for log(IL6) in the first- and second-stage regressions. Results are given in Table 7.3. We see that after adjusting for log(IL6) the causal estimate in each case has decreased (reflecting reduced weak instrument bias), its standard error has reduced (reflecting increased precision), and the F statistic has increased. With adjustment for a covariate, the relevant F statistic is a partial F statistic, representing the variation in the exposure explained by the IVs once the variation explained by the covariate has been accounted for. This is calculated from an analysis of variance (ANOVA) model.

7.6.3 Borrowing information across studies

The IV estimator would be unbiased if we knew the true values for the average exposure in different genetic subgroups. In a meta-analysis context [Thompson et al., 2005], we can combine the estimates of genotype–exposure association from different studies to give more precise estimates of exposure levels in each genetic subgroup. In the 2SLS method, an individual participant data (IPD) fixed-effect meta-analysis for data on individual i in study m with exposure x_{im}, outcome y_{im} and g_{ikm} for number of minor alleles (0, 1, or 2) of genetic

	Not adjusted		Adjusted	
IV estimate	Estimate (SE)	F statistic	Estimate (SE)	F statistic
Using rs1205	0.219 (0.201)	79.6	0.173 (0.196)	100.2
Using rs1417938	−0.457 (0.407)	27.6	−0.458 (0.362)	37.2
Using rs1800947	0.354 (0.325)	28.6	0.324 (0.316)	36.5
Using all 3 SNPs	0.186 (0.194)	24.4	0.127 (0.188)	32.2

TABLE 7.3
2SLS estimates and standard errors (SE) of the causal effect of log(CRP) on fibrinogen, and F statistic for regression of log(CRP) on IVs, calculated using each SNP separately and all SNPs together in per allele additive model, without and with adjustment for log(IL6) in the Cardiovascular Health Study.

variant k ($k = 1, 2, \ldots K_m$) is:

$$\begin{aligned} x_{im} &= \alpha_{0m} + \sum_{k=1}^{K_m} \alpha_{km}\, g_{ikm} + \varepsilon_{Xim} \\ y_{im} &= \beta_{0m} + \beta_1\, \hat{x}_{im} + \varepsilon_{Yim} \\ \varepsilon_{Xim} &\sim \mathcal{N}(0, \sigma_X^2); \varepsilon_{Yim} \sim \mathcal{N}(0, \sigma_Y^2) \text{ independently.} \end{aligned} \tag{7.6}$$

The exposure levels are regressed on the IVs using a per allele additive linear model separately in each study, and then the outcome levels are regressed on the fitted values of exposure ($\hat{x}_{im}$). The terms α_{0m} and β_{0m} are study-specific intercept terms. Here we assume homogeneity of variances across studies; we can use Bayesian methods to allow for possible heterogeneity (see Section 9.6).

If the same genetic variants are measured in each study and are assumed to have the same effect on the exposure, we can use common genetic effects (i.e. $\alpha_{km} = \alpha_k$) across studies by replacing the first line in equation (7.6) with:

$$x_{im} = \alpha_{0m} + \sum_{k=1}^{K} \alpha_k\, g_{ikm} + \varepsilon_{Xim} \tag{7.7}$$

If the assumption of common genetic effects is correct, this will improve the precision of the fitted values ($\hat{x}_{im}$) and reduce weak instrument bias.

To illustrate this, we consider the Copenhagen City Heart Study (CCHS), Edinburgh Artery Study (EAS), Health Professionals Follow-up Study (HPFS), Nurses Health Study (NHS), and Stockholm Heart Epidemiology Program (SHEEP), which are cohort studies or case-control studies measuring CRP and fibrinogen levels at baseline [CCGC, 2008]. In case-control studies, we use the data from controls alone since these better represent cross-sectional population studies. These five studies measured the same three SNPs on the *CRP* gene: rs1205, rs1130864 and rs3093077 (or rs3093064, which is in complete linkage disequilibrium with rs3093077). We estimate the causal effect

Study	N	F	df	Causal estimate (SE)	Observational estimate (SE)
CCHS	7999	29.6	(3, 7995)	−0.286 (0.373)	1.998 (0.030)
EAS	650	6.9	(3, 646)	0.754 (0.327)	1.115 (0.056)
HPFS	405	5.3	(3, 401)	0.758 (0.423)	1.048 (0.081)
NHS	385	6.1	(3, 381)	−0.906 (0.636)	0.562 (0.114)
SHEEP	1044	10.5	(3, 1040)	0.088 (0.345)	1.078 (0.051)
Different genetic effects		14.4	(15, 10463)	0.021 (0.195)	
Common genetic effects		56.6	(3, 10475)	−0.093 (0.225)	
Study-level estimates				0.234 (0.174)	

TABLE 7.4
Estimates of effect of log(CRP) on fibrinogen (μmol/l) from each of five studies separately and from meta-analysis of studies: number of participants (N), F statistic (F) with degrees of freedom (df) from per allele additive regression of exposure on three SNPs used as IVs, causal estimate using 2SLS with standard error (SE), observational estimate with SE. Fixed-effect meta-analyses conducted using individual-level data with different study-level genetic effects, common pooled genetic effects, and combining study-level estimates with inverse-variance weighting.

using the 2SLS method with different genetic effects (model 7.6), common genetic effects (model 7.7) and by a fixed-effect meta-analysis of estimates from each study.

Table 7.4 shows that the studies analysed separately have apparently disparate causal estimates with large SEs. The meta-analysis estimate assuming common genetic effects across studies is further from the confounded observational estimates and closer to the IV estimate from the largest study with the strongest instruments (CCHS) than the model with different genetic effects, suggesting that the latter suffers bias from weak instruments.

The pooled estimate from the study-level meta-analysis is greater than those from the individual-level meta-analyses. Although the CCHS study has about 8 times the number of participants as SHEEP and 12 times as many as EAS, its causal estimate has a larger standard error. The standard errors in the 2SLS method are known to be underestimated when the correlation due to confounding is strong, especially with weak instruments (Section 7.4.2) [Stock and Yogo, 2002]. Also, Figure 7.5 showed that causal estimates nearer to the observational association have lower variance. So a study-level meta-analysis may be biased due to overestimated weights in the studies with more biased estimates.

Returning to the example of data from the Copenhagen General Population Study considered in Section 7.2.1, if we use the IPD (model 7.6) to

Substudies	Meta-analysis	p-value	Different genetic effects	p-value	Common genetic effects	p-value
1	−0.05 (0.15)	0.76				
5	−0.01 (0.15)	0.95	−0.03 (0.15)	0.85	−0.05 (0.15)	0.75
10	0.09 (0.14)	0.54	0.04 (0.14)	0.80	−0.05 (0.15)	0.76
16	0.23 (0.14)	0.09	0.15 (0.14)	0.26	−0.05 (0.15)	0.75
40	0.46 (0.13)	< 0.001	0.30 (0.13)	0.02	−0.04 (0.15)	0.77
100	0.83 (0.11)	< 0.001	0.68 (0.11)	< 0.001	−0.04 (0.15)	0.77
250	1.27 (0.08)	< 0.001	1.15 (0.08)	< 0.001	−0.04 (0.15)	0.78

TABLE 7.5
2SLS estimates of causal effect (standard error) of log(CRP) on fibrinogen from the Copenhagen General Population Study divided randomly into substudies and combined: using fixed-effect meta-analysis of substudy estimates, and using individual patient data (IPD) with different or common genetic effects across substudies.

combine the substudies in the meta-analysis rather than combining estimates from each substudy, then the pooled estimates are somewhat less biased (Table 7.5). If we additionally assume common genetic effects across studies (model 7.7), then we recover close to the original estimate based on analysing the full dataset as one study: weak instrument bias has been eliminated.

7.7 Discussion

This chapter has demonstrated the effect of weak instrument bias on causal estimates in real and simulated data. The magnitude of this bias depends on the statistical strength of the association between instrument and exposure.

Weak instrument bias can reintroduce the problem that IVs were developed to solve. It is misleading not solely because it biases estimates, but because estimates suffering from the bias do not provide a valid test of the null hypothesis. Weak instruments may convince a researcher that an observational association that they have estimated is in fact causal. The reason for the bias is that the variation in the exposure explained by the IV is not large enough to dominate the variation in the exposure caused by chance correlation between the IV and confounders.

While the magnitude of the bias depends on the instrument strength through the expected or mean F statistic, for a study of fixed size and underlying instrument strength, an observed F statistic greater than its expected value corresponds to an estimate closer to the observational association with

greater precision; conversely an observed F statistic less than the expected value corresponds with an estimate further from the observational association with less precision. Simply relying on an F statistic from an individual study is over-simplistic and threshold rules such as ensuring $F > 10$ may cause more bias than they prevent.

7.7.1 Bias–variance trade-off

Using the 2SLS method, we demonstrated a bias–variance trade-off for the number of instruments used in IV estimation. For a fixed mean F statistic, as the number of instruments increases, the precision of the IV estimator increases, but the bias also increases. Using the LIML method, bias did not increase with the number of instruments, but the precision was slightly lower than for 2SLS. When using 2SLS, we seek parsimonious models of genetic association, for example using per allele additive models and including only IVs with a known association with the exposure, based on biological knowledge and external information. Provided the data are not severely misrepresented, these should provide the best estimates of the causal effect. Again, *post hoc* use of observed F statistics to choose between instruments may cause more bias than it prevents.

7.7.2 Combatting weak instrument bias in practice

Ideally, issues of weak instrument bias should be addressed prior to data collection, by specifying sample sizes, instruments, and genetic models using the best prior evidence available, to ensure that the expected values of F statistics are large. Where this is not possible, our advice would be to conduct sensitivity analyses using different IV methods, numbers of instruments and genetic models to investigate the impact of different assumptions on the causal estimate.

Testing the association between the outcome and each IV in turn (without estimating a causal effect) is a valid test of a causal relationship even with weak instruments. If there is a single IV, then an expected F statistic of 5 corresponds to a p-value in the regression of the exposure on the IV of around 0.03. It is perhaps unlikely that an IV would be considered for use in a dataset if the expected p-value were much greater than 0.03, and so bias from weak instruments would not be expected to be an issue in practice with a single IV. If there are multiple IVs, LIML or Bayesian methods could be used in the analysis, as the estimates from these are less biased than the 2SLS estimate. A difference between the 2SLS and LIML IV estimates is evidence of possible bias from weak instruments. The use of Fieller's theorem, the Anderson–Rubin test statistic or a Bayesian posterior distribution for inference is recommended.

It is also possible to summarize multiple SNPs into a single variable to reduce weak instrument bias using an allele score. Details about how to construct such a score are given in Chapter 8.

Adjustment for covariates helps reduce weak instrument bias. Including predictors of the exposure in the first-stage regression, or predictors of the outcome in the second-stage regression, also increases precision of the causal estimate. The former will also increase the F statistic for the IVs, and thus reduce weak instrument bias.

This chapter has considered bias in a one-sample Mendelian randomization setting. If the genetic associations with the exposure and outcome are estimated in non-overlapping sets of individuals, then bias from weak instruments will act in the direction of the null (Section 9.8.2). Although bias is never welcome, the direction of bias in a two-sample Mendelian randomization analysis means that a non-null causal effect estimate will not simply be an artefact of weak instrument bias.

7.7.3 Bias in study-level meta-analysis

In a meta-analysis context, bias is a more serious issue, as it arises not only from the bias in the individual studies, but also from the correlation between causal effect estimates and their variances which results in studies with effects closer to the observational estimate being over-weighted. By using a single IPD model, we can reduce the second source of bias. Additionally, we can pool information on the genetic association across studies to strengthen the instruments. The assumptions of homogeneity of variances and common genetic effects across studies made in Section 7.6.3 are overly restrictive in practice; more reasonable extensions of IV methods to a meta-analysis context are discussed in Chapter 9.

7.7.4 Caution about validity of IVs

Finally, we recall that the use of a genetic instrument in Mendelian randomization relies on certain assumptions. In this chapter we have assumed, although these may fail in finite samples, that they hold asymptotically. If these assumptions do not hold, for example if there were a true correlation between the instrument and a confounder, then IV estimates can be entirely misleading [Small and Rosenbaum, 2008].

7.8 Key points from chapter

- Bias from weak instruments can result in seriously misleading estimates of causal effects. Studies with instruments having large expected F statistics are less biased on average. However, if a study by chance has a larger

observed F statistic than expected, then the causal estimate will be more biased.

- Coverage levels with weak instruments can be poorly estimated by methods which rely on assumptions of asymptotic normality.
- Data-driven choice of instruments or analysis can exacerbate bias. In particular, any threshold guideline such as ensuring that an observed F statistic is greater than 10 is misleading. Methods, instruments, and data to be used should be specified prior to data analysis. Meta-analyses based on study-specific estimates of causal effect are susceptible to bias.
- Bias can be alleviated by use of measured covariates and parsimonious modelling of the genetic association (such as a per allele additive SNP model rather than one coefficient per genotype). This should be accompanied by sensitivity analyses to assess potential bias, for example from model misspecification.
- Bias can be reduced substantially by using LIML, Bayesian and allele score (see next chapter) methods rather than 2SLS, and bias in practice with a single IV should be minimal. Nominal coverage levels can be maintained by the use of Fieller's theorem with a single IV, and confidence intervals from the Anderson–Rubin test statistic or Bayesian MCMC methods with multiple IVs.