


Appendix 1: sample size requirements for instrumental variable analysis with genetic instruments

For given Type 1 and Type 2 error probabilities and study design, the sample size required to detect an effect of size $d$ is proportional to $1/d^2 V$, where $V$ is the Fisher information (expectation of minus the second derivative of the log-likelihood of the effect size) from a single observation. For a single observation from a logistic regression model (in which the effect is measured as the log OR), the Fisher information is $f(1/C_0 f) v$, where $f$ is the probability of being a case, and $v$ is the variance of the predictor variable.

For a cohort design testing for association of a rare disease ($f$ close to 0) with a quantitative trait that is scaled to have variance of 1, the Fisher information is simply the total number $n$ of cases yielded by the cohort study. For a case–control design with $N$ cases and $N$ controls ($f=0.5$), testing for an effect on disease risk of genotype (coded as 0, 1, 2) at an SNP with allele frequency $p$, the Fisher information is $2N \times 0.5 \times 0.5 \times 2p(1-p) = Np(1-p)$. For allele frequency 0.2, this evaluates to $N/6.25$. Thus, in this situation the number $N$ of cases required for a case–control study to detect the effect (measured as log OR associated with one extra copy of the disease-associated allele) is 6.25 times larger than the number $n$ of cases required for a cohort study to detect an effect of the same size (measured as log OR associated with change of 1 SD) of a continuous trait on disease risk.

In reality, the size of the genotypic effect $\gamma_g$ on the intermediate phenotype is usually modest: typically no more than 0.25 SD for each extra copy of the trait-raising allele. As sample size scales inversely with the square of the effect size, this implies that the case–control collection would have to be 100 ($16 \times 6.25$) times larger (in terms of number of cases) than the cohort study for the effect of genotype on disease to be detected in a conventional significance test. For a Bayesian hypothesis test, the sample size requirements for the case–control study of genotype–disease association are similar. Bayesian sample size requirements for an experiment comparing two hypotheses can be calculated by specifying the expected log-likelihood ratio (ELOD) favouring the true hypothesis over the alternative. For a given effect size $d$, the sample size required for 90% power to detect this effect at 5% significance in a classical hypothesis test is $(Z_{0.9} + Z_{0.975})^2/(d^2 V) = 10.5/(d^2 V)$, and the sample size required for an ELOD of log(100) favouring this effect size over the null is $2 \log_e(100)/(d^2 V) = 9.2/(d^2 V)$.

Commentary: To cause or not to cause confusion vs transparency with Mendelian Randomization

Els Goetghebeur

Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281, S9 B-9000 Gent, Belgium. E-mail: els.goetghebeur@ugent.be

Accepted 27 April 2010

As a natural experiment, random inheritance of alleles promises to allow unveiling causal effects of changeable traits. Fulfilling this warrants statistical research from several angles and we welcome a contribution on Bayesian methods for instrumental variables analysis with genetic instruments.1 Exploiting
the Bayesian paradigm, the authors emphasize features less valued in a more standard frequentist approach. We reflect on these relative to the estimating equations approach to instrumental variables (IV) analysis, which we first briefly outline.

Protecting the type I error has been of key concern to frequentists recognizing that detailed models for continuous covariates are imperfect. Semi-parametric methods for causal effects with IVs make as few assumptions as possible aiming for robustness: consistency of the causal null hypothesis test under just the fundamental IV assumptions.

For estimation, the causal parameter is defined directly as a contrast between expected outcome $Y$ and expectation of latent outcome $\hat{Y}(x_0)$, the potential outcome with exposure $X$ set at some reference level $'x_0'$. For binary outcomes, $\beta_{sc}$ may represent a linear shift on the logit scale:

$$\logit[E(Y|G,X)] - \beta_{sc}(X-x_0) = \logit [E(Y(x_0)|G,X)]$$

(1)

Unbiased estimating equations for $\beta_{sc}$ are then derived from zero correlation between backtransformed outcome $\hat{Y}(x_0) = \expit \{ \logit [E(Y|G,X)] - \beta_{sc}(X-x_0) \}$ and gene $G$, as implied by IV assumptions. With non-linearity, estimation requires further (logistic) modelling of the observed association $E(Y|G,X)$. Additional modelling improves efficiency without jeopardizing consistency under the null. Absent causal interaction between exposure and instrument, $E(Y|x_0)$ follows.

In contrast, the Bayesian approach defines $\beta_s$ indirectly as a parameter in a two-stage logistic regression association model. A causal interpretation is claimed under parametric constraints. Even then, it turns out, this association parameter need not equal a structural causal effect measure, hence biased estimation follows even in large samples. This approach—as its frequentist counterpart—is nevertheless popular because genotype–covariate and genotype–outcome associations may be measured in different studies.

The approach comes with further drawbacks and advantages.

(1) Lack of direct definition of the causal parameter disables a clean bias assessment when ultimately evaluating consequences of assumptions (breaking down) as generally advocated.

(2) Two novel targets for inference are introduced: (i) a causal effect evaluation via comparative testing; and (ii) a diagnostic test of no pleiotropy based on multiple instruments. While it is uncomfortable that test (ii) relies heavily on parametric assumptions, problems with (i) are more severe.

Rather than testing for a causal effect of $X$ or not, one focuses on $\theta$, the ratio of $\beta_s$ and crude association regression coefficient $\gamma_s$, and contrasts simple hypotheses: ‘there is no causal effect, $\theta=0$’ vs ‘there is no confounding, $\theta=1$’. This is a question of interest more to data analysts than clinically. When both a causal effect and confounding exist, it is anyone’s guess which will dominate the likelihood ratio. With such dubious test result, why not prefer reliable estimation/testing of $\beta_s$?

(3) Parametric assumptions involve the shape of the error distribution besides expected values: robustness properties are abandoned necessitating a sensitivity analysis. The power of Markov chain Monte Carlo (MCMC) is that posterior distributions can be obtained by generating the random parameter values under the proposed sequence of parametric models. In principle, these models can be quite complex as long as the sequence of conditional models does not lead to an excessive computational burden. Here we have:

(i) Linear regression of $X$ on $G$ with additive effects for the ordered levels of $G$. How well should residuals be estimated to guarantee reliable causal effect estimators from the second-stage model?

(ii) Logistic regression of risk on $E(X|G)$ with additive effect of $\varepsilon = X - E(X|G)$:

$$\logit(p_i) = \beta_0 + \beta_s E(X_i|G_i) + \beta_s \varepsilon_i$$

must be approximate at best. Logistic regression is ultimately abandoned in favour of probit regression to invoke so-called conjugate distributions, which yield posterior distributions of known distributional form and substantially simplify calculations.

(4) With all models in place, MCMC allows for finite sample inference on the joint parameter distribution. In ref., the frequentist solution for failing asymptotic inference with weak instruments produces confidence intervals with correct coverage, which seems all we need. Still, the flexibility of MCMC—when it works—is attractive, especially with logistic regression. Indeed, even when association equals causation, non-collapsibility of the logistic function means that regression coefficient $\beta_s$ depends on the covariates further conditioned on; here $E(X|G)$ and residual $\varepsilon$, a latent variable measured with error. This complicates interpretation and applicability of $\beta_s$. To obtain a causal parameter independent of the observed $X-G$ relationship, additional work and assumptions (no causal interaction with instrument) are needed to find, for instance, the population averaged risk difference. This is elegantly achieved not only using MCMC, but also with the frequentist approach.

(5) For generalized use in routine practice, causal models had indeed better allow for measurement error (on exposure $X$) and for missingness (at random). With corresponding conditional parametric models in place, the Bayesian MCMC
adaptation is straightforward. Today’s frequentist application of structural models offers no less under less stringent assumptions. A double instrumental variable approach incorporates not only random but also systematic measurement error without needing to specify details of the measurement error distribution. The suggestion that those methods cannot handle missingness, latent variables or measurement error is outdated.2,6–9

In summary, we agree that MCMC provides a powerful tool, allowing to ‘simply’ blend many parametric association models to arrive at a joint distribution of unknown parameters. For causal inference, however, the most important work lies in carefully laying out exactly what causal parameter is targeted and under which untestable assumptions it is consistently estimated—or what bias to expect. This seems poorly understood in this cross-sectional case, and will require much more work for repeated exposures, where nested structural modeling is needed if one wishes to incorporate exposure effects that interact with intermediate variables on the causal path from initial exposure to final outcome.10 Arriving at a parameter which conditions on a latent variable that may be population specific, or at a test of two simple hypotheses that are both wrong is felt to be unhelpful. We prefer estimating well-understood causal parameters directly. Considering model uncertainty, robustness properties are welcome and a sensitivity analysis should be common place. Simulations as in refs3,5 could reveal clearly how bias may emerge here even in the ‘ideal case’.

Funding
The author acknowledges support from the IAP (Interuniversity Attraction Pole) research network grant nr. P06/03 from the Belgian government (Belgian Science Policy).

Conflict of interest: None declared.

References