Appendix

Computing Models

The main text discussed a familiar linear model that might be used when the outcome variable is measured on a continuous scale. We continue to use that familiar model to discuss estimating the distribution of treatment effects. Then we expand the discussion to non-linear models.

Using that familiar linear model, to derive the posterior we need evidence for auxiliary parameter θ, but it is difficult to identify a source for that evidence. Here we follow the arguments in Hoff (Hoff, 2009, pp. 154-159) and with a slight modification suggested by Lancaster (Lancaster, 2004, p. 135).

Because we have no knowledge of θ, Hoff suggests using a semi-conjugate prior distribution. This term means that we depart from the pure Bayesian logic and allow the data to affect our belief about the prior. Hoff provides a suitable estimator, but we follow a similar, frequently used estimator attributed to Zellner (Zellner, 1986). Conveniently, the Stata software provides computing for this estimator.

Zellner assumes that β (from the prior distribution) is normally distribution with mean vector of μ and a covariance of . Following a specialized application of Zellner’s approach, the μ is assumed to equal a vector of zeros, and the V is:

[A1] 

Becauseis fixed, the assumption about V is just an assumption about g and σ2. The g is the “g-prior”. The best value for g is not apparent, but the logic is that if the data provide a great deal of information, then we want the prior to be weakly informative; hence V should be large, and consequently g should be large. In an illustration, Hoff suggests that g equal the data sample size; Lancaster suggests using the degrees of freedom from estimating the β. Of course, in large samples there is little difference between these two recommendations. We follow Lancaster.

Zellner’s assumptions lead to the posterior distribution of β conditional on σ2 as being normal with:

[A2] 

Note how closely these estimates look to ordinary least squares (OLS) estimates. For example, if the sample size were 1000, and if the model had 999 degrees of freedom, then the ratio g/(g+1) would be 999/1000 and practically the expectation for the posterior distribution of β would slightly shrink the OLS estimates. (Given the data model, the OLS estimates are the same as the likelihood estimates.) The posterior distribution foris written:

[A3] 

In equation [A3],  is the residual variance from an OLS regression and SSR is the sum of squares from the regression. Gam(…) denotes a two parameter gamma distribution. Here n=N-k where N is the sample size and k is the number of estimated parameters. It may be helpful to see that in this parameterization of the gamma distribution:

[A4] 

SSR equals n times the value of σ2 estimated from an OLS regression, so if n is large, the expected value of σ2 is close to SSR/n. This seems sensible. Also, note that as n gets large, the variance of 1/σ2 gets large – a desirable property because when the data are informative, we want the prior to have little role in the estimation.

We incorporate Zellner’s estimator into the fposterior(β,σ2) 🡪 G(…) 🡪 H(…) sequence. Although we could employ mechanical steps that do not use any Bayesian estimation software, for the sake of computing efficiency we use Stata’s bayesmh routine. Assuming a value of ρ, here are the steps:

1. Using known values of Ytreated and X in the estimation, compute an OLS regression to derive statistics required to prime Zellner’s estimator.
2. Using the statistics from the first step, apply Stata’s bayesmh with the Zellner g0 option (e.g. choose to set μ equal to zero) to derive a large number (K=10,000) of draws for βtreated and σtreated from the posterior distribution. These random draws are saved for use in step 4.
3. Steps 1 and 2 are repeated where Ycontrol is the outcome. Note that units who enter the treatment condition comprise the sample for the first regression and that units who enter the control condition comprise the sample for the second regression. This step leads to K draws of βcontrol and σcontrol from the posterior distribution. The random draws are saved for use in step 4.
4. The algorithm now enters K loops where K is the number of random draws of the parameters from step 2.
	1. For each of the K loops, impute values of Ytreat and Ycontrol when those values are missing. That is, when a unit is a member of the control group and the outcome is missing, impute an outcome conditional on the outcome in the treatment state, and when a unit is a member of the treatment group and the outcome is missing, impute an outcome conditional on the outcome under the control state. This step is the only time that an assumed ρ is used.
	2. For each of the K loops, compute the estimated treatment effect for each unit as δ=Ytreat-Ycontrol where one of those right-hand-side elements is observed and the other is imputed.
	3. For each of the K loops, compute a summary statistics τk based on the δ treatment effects.
5. Having estimated τ1 through τK, derive a credible interval to summarize the distribution.

Step 4a requires some additional statistical apparatus. It is essential to think about the distribution of etreated conditional on econtrol and the distribution of econtrol conditional on etreated. Normal theory (Bertsekas & Tsitsdklis, 2008, for example) says that, conditional on etreated, econtrol will be distributed as normal with mean equal to $\left({σ\_{control}}/{σ\_{treated}}\right)ρ\left(y\_{treated}-xβ\_{treated}\right)=$ $\left({σ\_{control}}/{σ\_{treated}}\right)ρe\_{treated} $and variance equal to$\left(1-ρ^{2}\right)σ\_{control}^{2}$. Likewise, normal theory says that etreated will be distributed as normal with mean equal to $\left({σ\_{treated}}/{σ\_{control}}\right)ρ\left(y\_{control}-xβ\_{control}\right)$ = $\left({σ\_{treated}}/{σ\_{control}}\right)ρe\_{control} $and variance equal to$\left(1-ρ^{2}\right)σ\_{treated}^{2}$. Thus, unless ρ = 0, knowledge of Ytreated will tell us something about the distribution of econtrol, and knowledge of Ycontrol will tell us something about the distribution of etreated. The algorithm takes these conditional distributions into account, explaining why results are sensitive to assumptions about ρ.

Several summary measures for the distribution of treatment effects are available. As suggested above, these might be the mean or the median, but those familiar statistics do not tell us much more about the distribution than we learn from traditional measures from RCTs. A more informative yet simple summary measure is the proportion of units who benefit from treatment, the proportion who are harmed by treatment, and the proportion for whom treatment has no effect. An alternative might be the proportion who benefit substantially, the proportion who benefit marginally, and so on. The analyst can be creative when describing the distribution of effects.

Although these five steps lead to a credible interval for τ, we have assumed knowledge of ρ, and in fact ρ is unknown but presumably bounded between 0 and 1. If we assume first that ρ = 0, then we might derive a credible interval for τ as. If we assumed second that ρ = 1, then we might derive a credible interval for τ as. The bounded interval is thenwhere τL is the lower of τL,ρ=0 and τL,ρ=1 and where τH is the higher of τH,ρ=0 and τH,ρ=1. The bounded interval is not a credible interval, but it is informative of the distribution of treatment effects.

For illustration, the discussion sbove assumes a continuous outcome measure modeled with a regression whose error terms are distributed as independent and identically normal. This model may not fit the investigator’s problem, but alternatives are available. We consider outcomes measured on a binary scale, outcomes that are ordered, and outcomes that are countable.

Outcome measures are often measured on a binary scale where 1 is considered better than 0. We can adapt [A1] to this role by treating Ytreated and Ycontrol as unobserved latent variables. That is, we do not observe Y; rather, we observe Z:

[A5] 

The L are unknown thresholds, which might be the same for both treated and control, although it is entirely possible that the β remain the same across the two equations and treatment works by shifting the thresholds.[[1]](#endnote-1) Typically the thresholds are assumed to be zero.

Estimation requires modification to steps 1-5 as outlined in the main text. Step 1 is inapplicable. In step 2, we choose a different likelihood (Stata’s Probit option) and a prior distribution (a flat prior). In step 3, we make posterior draws of the parameters β\* and L\*. The “\*” denotes that parameters are proportional to β and L – specifically β\*=β/σ and L\*=L/σ. This is the familiar but immaterial problem that the σ are not identified in a Probit model. When we simulated e1 and e2 for the linear model, we observed values for Y, so we sample values from a conditional distribution. For the binary problem, we only know when Y passes a threshold, so we sample from a censored distribution. A formula appears in (Maddala, 1983, p. 269). That is, unless ρ = 0, the imputed value for Ytreated will depend on whether Zcontrol = 0 or 1; likewise, unless ρ = 0, the imputed value for Ycontrol will depend on whether Ztreated = 0 or 1. Note that imputation of the Y’s is only an interim step; the final step is imputation of the Z’s.

Given binary outcomes, and considering the summary statistics τ, it may be useful to identify three post-treatment conditions: (1) Ztreatment = 1 and Zcontrol = 0, (2) Ztreatment = 0 and Zcontrol = 0 or Ztreatment = 1 and Zcontrol = 1, and (3) Ztreatment = 0 and Zcontrol = 1. In the first case, we might say that treatment is helpful; in the second case, we might say it is ineffective; and in the third case, we might say that treatment is harmful. The credible interval regards the proportion of units who benefited from treatment, the proportion for whom treatment did neither good nor harm, and the proportion for whom treatment was harmful.

As an illustration, we simulated data according to [A5] after setting L=0 for both treated and control units. In the simulation, ρ = 0.7 but of course this is presumed unknown at the time of estimation. About 25% of units benefit from treatment, about 72% receive no benefit, and about 3% are harmed. Assuming that ρ = 0, we estimate that 34% are benefited (31% to 37%) and few are harmed (7% to 10%). Assuming that ρ = 1, we estimate that 23% are benefited (18% to 28%) and few are harmed (0% to 0.5%). Clearly the credible intervals are sensitive to assumptions about ρ, but nevertheless, as evaluators we have learned a great deal: Although treatment is unhelpful for most units, regardless of the value of ρ a significant proportion benefit, and those who benefit from treatment greatly outnumber those who are harmed.[[2]](#endnote-2)

Outcome measures are sometimes measured on an ordinal scale. For example, when there are three possible outcomes, [A1] might be rewritten as:

[A6] 

An ordered Probit model is useful. A censored normal distribution again comes into play, but for the binary Probit, we could easily write an analytical statement for the censored normal distribution; this is difficult for the ordered Probit with an arbitrary number of outcome categories. Thus, instead of programming an analytical statement, we simulate data from the bivariate normal distribution to impute outcome categories for Ycontrol when outcome categories for Ztreated are known and to impute outcome categories for Ytreated when outcome categories for Zcontrol are known.

As an illustration, we simulated data. Unlike before, we set X to have a uniform distribution between -0.5 and 0.5, and we set thresholds of 0 and 1 for three outcomes. In the simulation, ρ = 0.7 but again this is presumed unknown at the time of estimation. About 37% of units benefit from treatment, about 55% receive no benefit, and about 9% are harmed. Assuming that ρ = 0, we estimate that 44% are benefited (41% to 47%) and that 22% are harmed (20% to 25%). Assuming that ρ = 1, we estimate that 31% are benefited (21% to 40%) and that fewer than 1% are harmed (0% to 0.1%). As before, the credible intervals are sensitive to assumptions about ρ, but nevertheless, as evaluators we have again learned a great deal: Although treatment is unhelpful for most units, those who benefit from treatment outnumber those who are harmed, and if we assume that ρ is large because outcome in the treated and untreated states are highly correlated for the same individual, the difference between those who are helped and those who are harmed seems large.

Outcomes might be measured as counts, in which case an analyst could entertain a negative binomial model that comprises a mixture of the Poisson and normal. For a discussion of this mixture model see (Cameron & Trivedi, 1998). Presumably a normal mixture is realistic and it is convenient because it allows us to again use a bivariate normal distribution. To clarify, the negative binomial has a likelihood:

 [A7] 

Hereis the normal density with mean zero and variance 1. C is the counts. Because the error terms are unobserved, standard estimation practice is to remove them by integration.

We adopt a different estimation method for this problem. Assuming an uninformative prior, the posterior is equivalent to the likelihood, and the estimation sequence changes from to just. We use Stata’s bayesmh estimator with the evaluator option to sample from the posterior. All the parameters except ρ can be sampled, so the problem is to sample etreated when the unit is a member of the control group and econtrol when the unit is a member of the treated group. In general, this is a difficult programming problem, but the difficulty is minor under two conditions: ρ = 0 and ρ = 1. Given that these two are the benchmark boundary conditions, programming solutions for just these two assumptions has merit.[[3]](#endnote-3)

As a demonstration, we set λ such that:

[A8] 

Although we have set in the simulation, assuming that the two σ are equal is unnecessary, and the estimation algorithm does not assume this equality. In reality (that is, in simulation reality), for 39% of the cases, there is an improvement, for 34% of the cases there is no change, and for 27% of the cases there is a decline. (This presumes that higher counts are better than lower counts.) Suppose we assume that ρ = 1. This is equivalent to assuming that etreated = econtrol. However, we do not assume that the σ are the same. Given this assumption, a sizable proportion of the study group suffers harm (0.21 to 0.28), a sizable proportion suffer neither harm nor benefit (0.33 to 0.41), and the largest proportion benefit (0.34 to 0.43). Alternatively, we might assume that ρ = 0, in which case there is no correlation between etreated and econtrol. We conclude that a small proportion of units are harmed (0.06 to 0.11), that a larger proportion benefit (0.14 to 0.21), and the majority neither benefit nor suffer harm (0.71 to 0.78). Under either assumption, the proportion of those who benefit does not really exceed the proportion who are harmed.

In fact, the way we are thinking about the negative binomial process may be problematic. In the negative binomial setting, even if we set both σ to 0 so that for everyone, there is no guarantee thatfor everyone, because the Poisson process remains inherently random. A better way of thinking about the distribution of effects may be to set. This provides a clearer picture of how treatment effects what is controllable and is analogous to our suggested approach for the linear model.

Exercising substantive knowledge, suppose we conclude that a one count change is immaterial, so we treat a change of two counts or more as a material change. Given this assumption, and the assumption that ρ = 1, a small proportion are harmed (0.04 to 0.08), most are unchanged materially (0.72 to 0.78), and fewer than one-quarter benefit (0.15 to 0.22).

1. The model does not identify both the constant and the threshold. Typically, the threshold is constrained to zero and the constant is allowed to vary. In that case, the statement that treatment may vary the threshold is really a statement that treatment varies the constant. The two effects are empirically indistinguishable. [↑](#endnote-ref-1)
2. The illustration assumes that the outcomes are truly distinct and meaningful, but it may be that the binary outcome is just an indicator and that we should be interested in the distribution of the latent variable Ys instead of the observed variable Zs. Alternatively, we might be interested in the distribution of just the linear part of the latent variable and ignore the random part. [↑](#endnote-ref-2)
3. Conceptually, the numerical solution requires integration over the unobserved e conditional on the observed counts. This is a difficult programming problem, which we approximate by translating the normal into a point process described by a density with equal mass as each point. That density changes to a conditional density based on the estimated values of Xβ and the observed values of C because given Xβ and C, some values of e are more likely than others. Sampling from that univariate conditional density is straightforward. We could extend this same logic to sampling from a conditional bivariate normal distribution, but the programming is considerably more difficult and computation time is increased. [↑](#endnote-ref-3)