CHAPTER 23

Causal inference using multilevel models

Causal inference using regression has an inherent multilevel structure—the data give comparisons between units, but the desired causal inferences are within units. Experimental designs such as pairing and blocking assign different treatments to different units within a group. Observational analyses such as pairing or panel study attempt to capture groups of similar observations with variation in treatment assignment within groups.

23.1 Multilevel aspects of data collection

Hierarchical analysis of a paired design

Section 9.3 describes an experiment applied to school classrooms with a paired design: within each grade, two classes were chosen within each of several schools, and each pair was randomized, with the treatment assigned to one class and the control assigned to the other. The appropriate analysis then controls for grade and pair.

Including pair indicators in the Electric Company experiment. As in Section 9.3, we perform a separate analysis for each grade, which could be thought of as a model including interactions of treatment with grade indicators. Within any grade, let n be the number of classes (recall that the treatment and measurements are at the classroom, not the student, level) and J be the number of pairs, which is $n/2$ in this case. (We use the general notation n, J rather than simply "hard-coding" $J = n/2$ so that our analysis can also be used for more general randomized block designs with arbitrary numbers of units within each block.)

The basic analysis has the form

$$\begin{aligned} y_i &\sim \text{N}(\alpha_{j[i]} + T_i\theta, \sigma_y^2), \text{ for } i = 1, \ldots, n \\ \alpha_j &\sim \text{N}(\mu_\alpha, \sigma_\alpha^2), \text{ for } j = 1, \ldots, J. \end{aligned}$$

By including the pair indicators, this model controls for all information used in the design.

Here is the Bugs code for this model, as fit to classrooms from a single grade, and using `pair[i]` to index the pair to which classroom i belongs:

Bugs code

```
model {
  for (i in 1:n){
    y[i] ~ dnorm (y.hat[i], tau.y)
    y.hat[i] <- a[pair[i]] + theta*treatment[i]
  }
  for (j in 1:n.pair){
    a[j] ~ dnorm (mu.a, tau.a)
  }
  theta ~ dnorm (0, .0001)
  tau.y <- pow(sigma.y, -2)
  sigma.y ~ dunif (0, 100)
```

```
    mu.a ~ dnorm (0, .0001)
    tau.a <- pow(sigma.a, -2)
    sigma.a ~ dunif (0, 100)
    }
  }
```

Fitting all four grades at once. We can fit the above model to each grade separately, or we can expand it by allowing each of the parameters to vary by grade. For convenience, we use two index variables: `grade`, which indexes the grades of the classrooms, and `grade.pair`, which indexes the grades of the pairs. (This particular experiment involves $n = 192$ classrooms clustered into $J = 96$ pairs, so `grade` is a vector of length 192 and `grade.pair` has length 96. The entries of both vectors are 1's, 2's, 3's, and 4's.) Here is the Bugs model:

Bugs code

```
model {
  for (i in 1:n){
    y[i] ~ dnorm (y.hat[i], tau.y[grade[i]])
    y.hat[i] <- a[pair[i]] + theta[grade[i]]*treatment[i]
  }
  for (j in 1:n.pair){
    a[j] ~ dnorm (mu.a[grade.pair[j]], tau.a[grade.pair[j]])
  }
  for (k in 1:n.grade){
    theta[k] ~ dnorm (0, .0001)
    tau.y[k] <- pow(sigma.y[k], -2)
    sigma.y[k] ~ dunif (0, 100)
    mu.a[k] ~ dnorm (0, .0001)
    tau.a[k] <- pow(sigma.a[k], -2)
    sigma.a[k] ~ dunif (0, 100)
  }
}
```

Writing the model this way has the advantage that the Bugs output (not shown here) displays inferences for all the parameters at once. The treatment effects θ are large for the first two grades and closer to zero for grades 3 and 4; the intercepts α_j are highly variable for the first 11 pairs (which correspond to classes in grade 1), vary somewhat for the next bunch of pairs (which are grade 2 classes), and vary little for the classes in higher grades. The residual data variance and the between-pair variance both decrease for the higher grades, all of which are consistent with the compression of the scores for higher grades at the upper end of the range of data (see Figure 9.4 on page 174).

Controlling for pair indicators and pre-test score. As discussed in Section 9.3, the next step is to include pre-test class scores as an input in the regression. The treatments are assigned randomly within each grade and pair, so it is not necessary to include other pre-treatment information, but adding pre-test to the analysis can improve the precision of the estimated treatment effects.

For simplicity, we return to the model on page 503 that fits one grade at a time. We need to alter this Bugs model only slightly, by simply adding the term `+ b.pre.test*pre.test[i]` to the expression for `y.hat[i]`, and then at the end of the model, placing the prior distribution, `b.pre.test ~ dnorm(0,.0001)`.

Figure 23.1 displays the estimated treatment effects for the models controlling for pairing, with and without pre-test score. The final analysis, including pairing and pre-test as inputs, clearly shows a positive treatment effect in all grades, with narrower intervals than the corresponding estimates without including the pairing information (see Figure 9.5 on page 176).

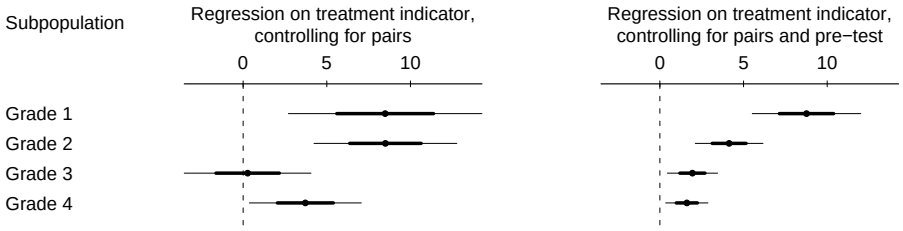

Figure 23.1 *Estimates, 50%, and 95% intervals for the effect of the Electric Company television show in each grade as estimated from hierarchical models that account for the pairing of classrooms in the experimental design. Displayed are regression coefficients for the treatment indicator: (a) also controlling for pair indicators, (b) also controlling for pair indicators and pre-test scores. Compare to Figure 9.5 on page 176, which shows estimates not including the pairing information. The most precise estimates are those that control for both pairing and pre-test.*

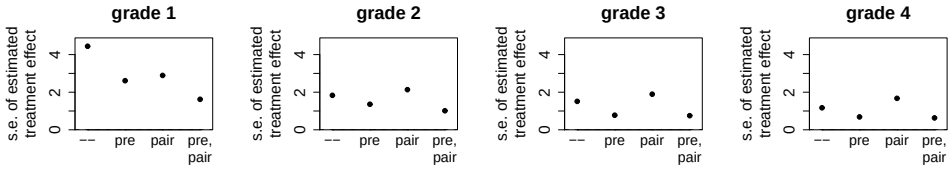

Figure 23.2 *Standard errors for estimated treatment effects in each grade for each of four models: (−−) with no other predictors in the model, (pre) controlling for pre-test, (pair) controlling for indicators for pairing, (pre, pair) controlling for pre-test and pair indicators. Unsurprisingly, controlling for more pre-test information tends to reduce estimation errors. Figures 9.5 and 23.1 display the estimated treatment effects and uncertainty bounds for the four models in each grade.*

To show more clearly the improvements from including more pre-treatment data in the model, Figure 23.2 displays the standard deviations of the estimated treatment effect in each grade as estimated from models excluding or including the pairing and pre-test information. In this case, the pre-test appears to contain more information than the pairing, and it is most effective to include both inputs, especially in grade 1, where there appears to be wide variation among classes.

Hierarchical analysis of randomized-block and other structured designs

More generally, pre-treatment information used in the design should be included as inputs so that the coefficients for the treatment indicators correspond to causal effects. For example, consider designs such as randomized blocks (in which data are partitioned into groups or "blocks," with random treatment assignment within each block) or latin squares (in which experimental units in a row × column data structure are assigned treatments randomly with constraints on the randomization so that treatment assignments are balanced within each row and column), the blocks, or the rows and columns, represent pre-treatment factors that can be accounted for using a multilevel model. This combines our general advice for causal inference in Chapter 9 with the idea from Chapter 12 of multilevel modeling as a general approach for handling categorical variables as regression predictors.

23.2 Estimating treatment effects in a multilevel observational study

In Section 10.1 we introduced the example of estimating the effect of the Infant Health and Development Program in the context of a (constructed) observational study. In that classical regression analysis, we included indicator variables for the 8 sites in order to control for unobserved site-specific characteristics that might be associated with both selection into treatment as well as the outcome (test scores). Here we extend to a multilevel model.

Varying intercepts

The simplest extension is just to allow for varying intercepts across site. If we denote post-treatment outcomes by y, treatment assignment by T, and the vector of confounding covariates for person i as X_i, we can write

$$\begin{aligned} y_i &\sim \text{N}(\alpha_{j[i]} + T_i\theta + X_i\beta,\, \sigma_y^2), \text{ for } i = 1, \ldots, n, \\ \alpha_j &\sim \text{N}(\mu_\alpha, \sigma_\alpha^2), \text{ for } j = 1, \ldots, J. \end{aligned}$$

Here, the matrix of predictors X does not include a constant term, since we have included a constant in the model for the α_j's. The treatment effect θ is estimated at 9.1 with a standard error of 2.1. (By comparison, our estimate from the classical regression with site indicators was 8.8 with standard error of 2.1, and the experimental benchmark is 7.4.)

Assumptions satisfied? The original justification for including site information when estimating the causal effect of the treatment was that it is plausible that unobserved site characteristics may be associated with both selection into treatment and subsequent test scores. However, the model above implicitly assumes that the intercepts (which capture unobserved characteristics of the sites) are independent of the other predictors in the model.

How can we resolve this conceptual inconsistency? One approach is to allow the intercepts to be correlated with the treatment variable. We can accomplish this by creating an aggregated version T^{agg} of the treatment variable, defined so that T_j^{agg} is the average value of T_i for the members of group j—in this case, the proportion who received the treatment, among the people in the dataset in site j. We then add this measure as a group-level predictor:

$$\alpha_j \sim \text{N}(\gamma_0 + \gamma_1 T_j^{\text{agg}}, \sigma_\alpha^2).$$

In our example, this changes the estimate of the treatment effect to 9.0, with the standard error virtually unchanged at 2.1. Addition of aggregated measures of the other predictors brings the estimate to 8.9.

Varying treatment effects

The demographic composition of the sample varied across sites, as did treatment implementation to some degree. Therefore we might expect treatment effects to vary also. A perhaps more interesting use of multilevel models in this example is to investigate variation in treatment effects across sites: we fit the model

$$\begin{aligned} y_i &\sim \text{N}\left(\alpha_{j(i)} + T_i\theta_{j(i)} + X_i\beta,\, \sigma_y^2\right), \text{ for } i = 1, \ldots, n, \\ \begin{pmatrix} \alpha_j \\ \theta_j \end{pmatrix} &\sim \text{N}\left(\begin{pmatrix} \mu_\alpha \\ \mu_\theta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\theta \\ \rho\sigma_\alpha\sigma_\theta & \sigma_\theta^2 \end{pmatrix}\right), \text{ for } j = 1, \ldots, J. \end{aligned}$$

Figure 23.3 displays the inferences for the treatment effects θ_j for each site from

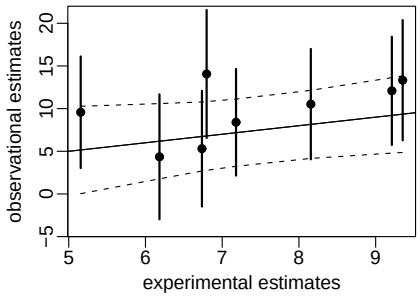

Figure 23.3 *Comparison of observational treatment effects (dark dots) and 95% intervals (dark vertical lines) across sites from multilevel models against the experimental benchmark (solid line with slope 1) and corresponding 95% confidence band (dotted curves). All observational intervals cover the corresponding experimental estimates; they all reflect greater uncertainty as compared to the experimental intervals as well.*

the multilevel model fit to observational data. The experimental estimates for each site (also calculated using a multilevel model) are referenced by the line with slope 1. A 95% interval for each observational estimate has been plotted (dark vertical line) for each site. The dotted curves display an approximate 95% confidence band for the experimental estimates. The observational intervals all cover the corresponding experimental estimate (our best comparison point for this example), but with greater uncertainty than the experimental estimates.[1]

23.3 Treatments applied at different levels

As noted in Section 20.1, treatments are sometimes implemented on *groups* of individuals (experimental units) rather than the individual units themselves. The choice of experimental design may be motivated by several concerns.

Suppose, for instance, that we want to evaluate a new method for teaching long division by teaching a random sample of third-grade students and then evaluating outcomes on math tests six months later. Sampling and logistical considerations would motivate a nested design such as including students in 40 classrooms across 20 schools in the study. It is easier and cheaper to train half the teachers (20 rather than 40) to implement the new method. Moreover, even if all teachers were trained, it would be inconvenient to divide each classroom into two parts, teaching the old method to half the students and the new method to the others. Finally if some of the students in a classroom were taught the new method and others the old, it is possible that the students would share information about the methods they were taught. This could influence their subsequent test scores and violate the assumption of independent outcomes which is standard in regression modeling (for further discussion of this principle of stable unit treatment values, see the end of Section 9.3). These are all motivations for randomizing classrooms rather than the students within classrooms (they might even motivate randomizing schools rather than classrooms). This is called a group- or cluster-randomized experimental design.

As with many of the other sampling and experimental designs discussed in this book, it is appropriate to analyze data from a grouped experiment using multilevel

[1] The displayed inferences actually come from a simpler version of the model in which the correlation ρ of the group-level errors was set to zero. For our example, the estimation with the full model including ρ was unstable, and there was no evidence that the correlation differed from zero.

models. In an extreme sense, it is as if we have only randomized J experimental units (where J is the number of groups) rather than the n individual units.

Analysis of an educational-subsidy program

In 1997 the Mexican federal government implemented Progresa (Programa de Educacion, Salud y Alimentacion), a program that provides cash subsidies to low-income families if they send their children to school (rather than, for instance, having them leave school to go to work at an early age) and visit health clinics. This program was randomly assigned to 506 eligible localities.[2]

Here we analyze a convenience subsample of the Progresa data that includes 81 localities and approximately 7000 households. The primary goal for this analysis is to determine if there was an effect of the program on enrollment in school.

A standard analysis that ignores the grouping might be to simply run a logistic regression of post-program enrollment on the treatment indicator and possibly some additional predictors to increase efficiency (baseline enrollment, work status, age, sex, and poverty status). This analysis yields an estimated treatment effect of 0.51 with standard deviation 0.09—a highly statistically significant result. This estimate implies that program availability increased the probability of enrollment by (at most) about 13%.

In contrast, we can build a multilevel logistic regression model of the form,

$$\begin{aligned} \Pr(y_i = 1) &= \text{logit}^{-1}(\alpha_{j[i]} + X_i\beta), \text{ for } i = 1, \ldots, n \\ \alpha_j &\sim \text{N}(U_j\gamma, \sigma_\alpha^2), \text{ for } j = 1, \ldots, J, \end{aligned}$$

where X is the matrix of individual-level predictors just described, and U is the matrix of group-level predictors, in this case simply a constant term and the treatment indicator.

Here is the corresponding Bugs model:[3]

Bugs code

```
model {
  for (i in 1:n){
    y[i] ~ dbin (p.bound[i], 1)
    p.bound[i] <- max(0, min(1, p[i]))
    logit(p[i]) <- Xbeta[i]
    Xbeta[i] <- a[village[i]] + b.1*enroll97[i] + b.2*work97[i] +
                b.3*poor[i] + b.4*male[i] + b.5*age97[i]
  }
  b.1 ~ dnorm (0, .0001)
  b.2 ~ dnorm (0, .0001)
  b.3 ~ dnorm (0, .0001)
  b.4 ~ dnorm (0, .0001)
  b.5 ~ dnorm (0, .0001)
  for (j in 1:J){
    a[j] ~ dnorm (a.hat[j], tau.a)
    a.hat[j] <- g.0 + g.1*program[j]
  }
  g.0 ~ dnorm (0, .0001)
  g.1 ~ dnorm (0, .0001)
```

[2] Localities not assigned to receive the program immediately were given the program a few years later.

[3] An alternative parameterization would store X as a matrix and express the linear predictor as `Xbeta[i] <- a[village[i]] + inprod(b[],X[i,])`, which would allow us to model the coefficients `b` more conveniently.

```
  tau.a <- pow(sigma.a, -2)
  sigma.a ~ dunif (0, 100)
}
```

The multilevel analysis yields a treatment coefficient estimate of 0.17 with a standard error of 0.20. In this case, correctly accounting for our uncertainty has a substantial impact on the results.[4]

Unmodeled varying intercepts? One strategy sometimes used (inappropriately) to account for the type of clustering observed in this experiment is to include indicator variables for each group—here each village. In this example such a model would yield a treatment effect estimate of 0.6 with standard error 1.4. While this estimate is also statistically insignificant, the uncertainty here is more than six times the uncertainty in the varying-intercept multilevel model.

A split-plot experiment

Figure 22.7 on page 499 shows the analysis of variance for a split-plot experiment: a design in which one set of treatments is applied at the individual level, and another set of treatments is applied at the group level. This is a randomized experiment with the probabilities of treatment assignment based on observed pre-treatment variables, and so the analysis is straightforward:

- Regress the outcome on the treatment indicators (these are treatments 1, 2 at the subplot (individual) level, A, B, C, D, E at the main-plot (group) level, and the 10 interactions A*1, A*2, ..., E*1, E*2.
- Also include in the regression the indicators for the 5 rows, the 5 columns, and the 25 main plots (groups).

For causal inference we should look at the coefficient estimates, which we display in Figure 23.4.

23.4 Instrumental variables and multilevel modeling

Section 10.5 discussed how instrumental variables can be used to identify causal effects in certain prescribed situations (when a valid instrument exists). We return to this same example, the randomized Sesame Street experiment, to illustrate extensions of the standard model to a multilevel framework.

The basic model can be written as

$$\begin{pmatrix} y_i \\ T_i \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} \alpha + \beta T_i \\ \gamma + \delta z_i \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \rho\sigma_y\sigma_T \\ \rho\sigma_y\sigma_T & \sigma_T^2 \end{pmatrix}\right), \text{ for } i = 1, \ldots, n,$$

where in this example z represents the randomized encouragement to watch Sesame Street, T represents whether the child subsequently watched or not (the desired "treatment" variable which in other contexts might be called the "compliance" variable), and y is the outcome measure, a post-treatment score on a letter recognition test. In this model, β is the causal effect of watching the show.

Recall that this experiment was randomized within combinations of site and setting. Therefore we first extend to include varying intercepts for each equation,

[4] We are analyzing here a nonrandom subset of the data and do not intend the results of this analysis to represent effects of the treatment over the entire experiment.

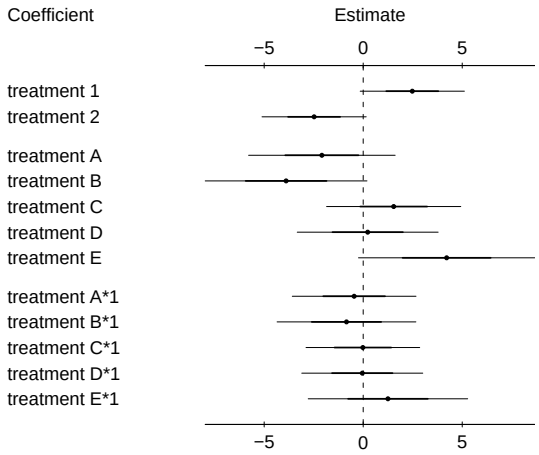

Figure 23.4 *Coefficient estimates for a split-plot experiment: an example of causal inference for a design with treatments at individual and group levels. The ANOVA table summarizing this analysis appears in Figure 22.7 on page 499. (This plot omits the five interactions* `A*2`*, . . . ,* `E*2`*, since, with only two levels of the numerical factor, these are simply the opposites of* `A*1`*, . . . ,* `E*1`*.)*

and we allow those intercepts to be correlated:

$$\begin{pmatrix} y_i \\ T_i \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} \alpha_{j[i]} + \beta T_i \\ \gamma_{j[i]} + \delta z_i \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \rho_{yT}\sigma_y\sigma_T \\ \rho\sigma_y\sigma_T & \sigma_T^2 \end{pmatrix}\right), \text{ for } i = 1, \ldots, n,$$
$$\begin{pmatrix} \alpha_j \\ \gamma_j \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} \mu_\alpha \\ \mu_\gamma \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho_{\alpha\gamma}\sigma_\alpha\sigma_\gamma \\ \rho\sigma_\alpha\sigma_\gamma & \sigma_\gamma^2 \end{pmatrix}\right), \text{ for } j = 1, \ldots, J. \qquad (23.1)$$

After creating a variable called `siteset` that represents the 9 existing combinations of site and setting (school or home), and bundling the outcome y and encouragement T into a single $n \times 2$ data matrix `yt`, we fit the following Bugs model:

Bugs code

```
model {
  for (i in 1:n){
    yt[i,1:2] ~ dmnorm (yt.hat[i,], Tau.yt[,])         # data model
    yt.hat[i,1] <- a[siteset[i]] + b*yt[i,2]
    yt.hat[i,2] <- g[siteset[i]] + d*z[i]
  }
  for (j in 1:J){
    ag[j,1:2] ~ dmnorm (mu.ag[1:2], Tau.ag[1:2,1:2])
    a[j] <- ag[j,1]
    g[j] <- ag[j,2]
  }

# data level
  Tau.yt[1:2,1:2] <- inverse(Sigma.yt[,])
  Sigma.yt[1,1] <- pow(sigma.y,2)
  sigma.y ~ dunif (0, 100)              # noninformative prior on sigma.a
  Sigma.yt[2,2] <- pow(sigma.t,2)
  sigma.t ~ dunif (0, 100)              # noninformative prior on sigma.b
  Sigma.yt[1,2] <- rho.yt*sigma.y*sigma.t
  Sigma.yt[2,1] <- Sigma[1,2]           # noninformative prior on rho
  rho.yt ~ dunif(-1,1)
```

```
  d ~ dnorm (0, .0001)
  b ~ dnorm (0, .0001)

# group level
  Tau.ag[1:2,1:2] <- inverse(Sigma.ag[,])
  Sigma[1,1] <- pow(sigma.a,2)
  sigma.a ~ dunif (0, 100)
  Sigma[2,2] <- pow(sigma.g,2)
  sigma.g ~ dunif (0, 100)
  Sigma[1,2] <- rho.ag*sigma.a*sigma.g
  Sigma[2,1] <- Sigma[1,2]
  rho.ag ~ dunif(-1,1)

  mu.ag[1] ~ dnorm(0, .0001)
  mu.ag[2] ~ dnorm(0, .0001)
}
```

The causal parameter of interest is `b`.

One advantage of Bugs is that it allows us to model the standard form of this model directly, but it turns out that modeling the reduced form of the model is more efficient and the algorithm will converge much more quickly. This just requires a simple change to the fourth line of the above model:

```
yt.hat[i,1] <- a[siteset[i]] + b*d*z[i]
```
Bugs code

Conditioning on pre-treatment variables

The instrumental variables model can be augmented by conditioning on pre-treatment scores as well, changing the fourth and fifth lines of the model to:

```
yt.hat[i,1] <- a[siteset[i]] + b*d*z[i] + phi.y*pretest[i]
yt.hat[i,2] <- g[siteset[i]] + d*z[i] + phi.t*pretest[i]
```
Bugs code

and specifying prior distributions for `phi.y` and `phi.t`.

The results from this model indicate a treatment effect distribution centered at 13.5 with a standard error of about 3.8. Results from two-stage least squares were similar with an estimate of 14.1 and standard error of 3.9, although in general the two approaches can give different answers.

Varying treatment effects

Because randomization occurred at the individual level, we could also extend this model to include varying treatment effects (similar to the coding for the varying intercepts). In this example, however, the sample sizes in each group were too small to estimate varying treatment effects reliably.

Group-level randomization

It is common to see an instrument that was assigned at the group level. This occurs in the case of a group-randomized experiment or, for example, when state policies are used as instruments in an analysis of individual outcomes. In these settings, the varying intercept model presented here is appropriate; however, varying treatment effects cannot be identified.

23.5 Bibliographic note

The books and articles referred to in Sections 9.9 and 10.8 include many examples with treatments at different levels, but we have few specific references for causal inference and multilevel models. Oakes (2004) and the accompanying discussion consider the challenges of interpreting multilevel coefficients causally, and Sobel (2006) considers the assumptions involved in estimating treatment effects in the presence of interactions between individuals and groups.

For discussion of prior distributions for Bayesian instrumental variables models, see, for instance, Dreze (1976), Maddala (1976), Kleibergen and Zivot (2003), and Hoogerheide, Kleibergen, and van Dijk (2006).

23.6 Exercises

1. Fit a varying-intercept model (without varying slopes) to the Sesame Street data (in folder `sesame`) and compare to the results in Section 23.4.
2. Simulate data from a group randomized experiment and then fit using a classical and then a multilevel model. The confidence intervals from the classical fit should be too narrow.
3. Generate data from an instrumental variables model randomized at the group level where differences exist between groups. Fit a varying-intercept model to these data. Compare these results to results from a classical two-stage least squares fit as described in Section 10.5.
4. Generate data from an instrumental variables model randomized at the individual level where treatment effects vary across groups. Fit a varying-intercept, varying-slope model to these data. Compare these results to results from a classical two-stage least squares model.
5. Explain why varying treatment effects can be identified when the instrument is randomized at the individual level but not when the instrument is randomized at the group level.