Estimating the impact of air pollution using small area estimation

Male chimpanzee aggression towards females: a test of the sexual coercion hypothesis

mine cetinkaya-rundel

statistical science
Estimating the impact of air pollution using small area estimation
Figure 1.1: Geographical distribution of data from birth certificates, EPOS, and ECHOS.
Figure 1.2: Zip code areas that are merged to correct for areas with too few observations.
Figure 1.3: Distribution of observed rates of outcomes of interest

- Wheezing
- Wheezing in the last 12 months
- Doctor diagnosed ear infections
- Wheezing or asthma
- Doctor diagnosed asthma
- Medication use for wheezing or asthma
Figure 1.4: Distribution of air pollution exposure estimates.
Develop a model that makes use of information readily available in public data bases instead of having to collect data via costly follow-up surveys
Figure 1.3: Distribution of observed rates of outcomes of interest

- Rare outcomes + low sample sizes from each area
- Fluctuations highlighted might be due to chance, not real changes in the underlying risk factors
- Smoothing
\[ y_i | x_i \sim Poisson(\mu_i) \]
\[ \mu_i = E_i \theta_i \]
\[ E_i = n_i r \quad n_i \quad \text{sample size in each small area} \]
\[ r = \frac{\sum_{i=1}^{m} y_i}{\sum_{i=1}^{m} n_i} \quad \text{overall success rate} \]
\[ g(\theta_i) = \beta X_i + u_i + v_i + \alpha \]
\[ \beta X_i \quad \text{covariates} \]
\[ v_i \quad u_i \quad \text{area-level error} \]
\[ \alpha \quad \text{unit-level error} \]
\[ u_i \sim N(0, \sigma_u^2) \]
\[ v_i | v_{-i} \sim N \left( \frac{\sum_{j \sim i} v_j}{n_i}, \frac{\sigma_v^2}{n_i} \right) \]
\[ \text{conditional auto-regressive} \]
model fitting

WinBugs and R2WinBUGS

Gibbs sampler with four independent chains

100,000 iterations (20,000 burn-in iterations)

Uninformative priors assumed on all $\beta$s, $\pi(\beta) \sim N(0,1e5)$
models

air pollution variables: NO, NO$_2$, O$_3$, and CO

M1: air pollution + race/ethnicity

   no models with only air pollution exposure metrics since in public health literature it is believed that there is bound to be some confounding due to demographic variables

M2: air pollution + race/ethnicity
    + education + prenatal pay

M3: air pollution + race/ethnicity
    + education + prenatal pay
    + average per capita income per small area
Wheezing

<table>
<thead>
<tr>
<th>Observed/Expected</th>
<th>Model 1: Race only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>TOO FLAT</strong></td>
</tr>
<tr>
<td>Model 2: M1 + edu + prenatal pay</td>
<td><strong>BETTER</strong></td>
</tr>
<tr>
<td>Model 3: M2 + imputed income</td>
<td><strong>NO IMPROVEMENT</strong></td>
</tr>
</tbody>
</table>

ratio of obs / exp

Figure 4.2: Wheezing, with NO, NO, and O.
Doctor diagnosed ear infections, mean

<table>
<thead>
<tr>
<th>Observed/Expected</th>
<th>Model 1: Race only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NOT SO RARE OUTCOME</td>
</tr>
<tr>
<td></td>
<td>SMOOTHER</td>
</tr>
</tbody>
</table>

Model 2: M1 + edu + prenatal pay

Model 3: M2 + imputed income

Figure 4.12: Doctor diagnosed ear infections, with NO and CO

80 SMOKER

SMOOTHER

NOT SO RARE

OUTCOME

80

IMPROVEMENT

NOT

SO

RARE

OUTCOME

SMOOTHER

NO IMPROVEMENT

NO IMPROVEMENT

1.97

1.38

0.96

0.67

0.47

0.00
Male chimpanzee aggression towards females: a test of the sexual coercion hypothesis
Estimating the impact of air pollution using small area estimation data.
Estimating the impact of air pollution using small area estimation data

cases: paired interactions (dyads) between male and female chimpanzees

explanatory:
- aggression rates, copulation, offspring and paternity
- demographic information on the chimpanzees (female age, male rank in the population, etc.)
- swollen period (exposure time)

response:
- number of copulations in the reproductive window
- whether the dyad achieved a paternity in the conception window in question (0/1)
Estimating the impact of air pollution using small area estimation.

One of the response variables of interest is copulation counts. Figure 7 displays a histogram of these counts, as well as the empirical density and the Poisson density with the sample mean as the parameter. The observed copulation counts are clearly overdispersed when compared to the Poisson distribution. This is verified by the large discrepancy between the sample mean 4.754 and the sample variance 44.726. This suggests that distributions that allow for overdispersion of count data, such as quasi-Poisson or negative binomial, may be more appropriate for modeling these data than the obvious first choice of Poisson distribution.

The relationships between the response variable and the covariates were also investigated. As we can see in Figure 8, the distribution of copulation rates varies a lot depending on the female, but not as much depending on the male rank or parity status. The relationship between the copulation counts and the aggression rates in Figure 9 has a small positive correlation, but it does not indicate a strong trend neither by female, male rank or parity. The same behavior is observed for non-swollen aggression rates in Figure 10, male rank scores in Figure 11, relatedness in Figure 12, and female age in Figure 13.

Histograph of the copulations counts:

- Empirical density
- Poisson density

Poisson
Quasi-Poisson
Negative Binomial
Poisson with random effect for individual
Estimating the impact of air pollution using small area estimation.

\[
\text{mod.pois} = \text{glmer(} \text{cop_count ~ Parity} + \text{Swollen_agg_Z} + \text{NonSwol_agg_Z} + \text{M_Rank_JTFMDS}
\]
\[
+ \text{Rank_pd_Despot_ratio} + \text{Relatedness} + \text{fem_age_c} \quad \# \text{main effects}
\]
\[
+ \text{Parity*Swollen_agg_Z} + \text{Parity*NonSwol_agg_Z} \quad \# \text{interactions}
\]
\[
+ \text{Parity*M_Rank_JTFMDS} + \text{Parity*Relatedness}
\]
\[
+ \text{Swollen_agg_Z*M_Rank_JTFMDS} + \text{Swollen_agg_Z*Relatedness}
\]
\[
+ \text{Swollen_agg_Z*fem_age_c} + \text{NonSwol_agg_Z*M_Rank_JTFMDS}
\]
\[
+ \text{NonSwol_agg_Z*Relatedness} + \text{NonSwol_agg_Z*fem_age_c}
\]
\[
+ \text{M_Rank_JTFMDS*Rank_pd_Despot_ratio} + \text{M_Rank_JTFMDS*fem_age_c}
\]
\[
+ (1|\text{FEMALE}) \quad \# \text{random effects by female}
\]
\[
+ (1|\text{obs}) \quad \# \text{random effect for obs ID as a fix for overdispersion}
\]
\[
+ \text{offset(} \log(\text{Num_int_swollen})\text{)} \quad \# \text{offset}
\]
\[
data = \text{chimp}, \quad \text{family = "poisson"}\) \quad \# \text{poisson model}
\]

\[
\text{all.mod.pois} = \text{dredge(} \text{mod.pois, rank="AIC"}, \text{fixed=~offset(} \log(\text{Num_int_swollen})\text{)},
\]
\[
\text{trace=T, m.min=1) \quad \# \text{model selection}
\]
mod.bin = lmer(Paternity ~ Parity + Swollen_agg_Z + NonSwol_agg_Z + M_Rank_JTFMDS + Rank_pd_Despot_ratio + Relatedness + fem_age_c  # main effects
+ Parity*Swollen_agg_Z + Parity*NonSwol_agg_Z  # interactions
+ Parity*M_Rank_JTFMDS + Parity*Relatedness
+ Swollen_agg_Z*M_Rank_JTFMDS + Swollen_agg_Z*Relatedness
+ Swollen_agg_Z*fem_age_c + NonSwol_agg_Z*M_Rank_JTFMDS
+ NonSwol_agg_Z*Relatedness + NonSwol_agg_Z*fem_age_c
+ M_Rank_JTFMDS*Rank_pd_Despot_ratio + M_Rank_JTFMDS*fem_age_c
+ (1|FEMALE),  # random effects by female
data = chimp, family = binomial)  # logistic model
all.mod.bin = dredge(mod.bin, rank="AIC", trace=T, m.min=1)  # model selection
2011 (UCLA): LAPD 911 calls

2012 (Duke): kiva.com

2013 (Duke + UNC + NCSU): eHarmony

2014 (Duke + UNC + NCSU): GridPoint
cases: 1M matches generated by the eHarmony algorithm

success: user (female) or candidate (male) sends message within 7 days of match

variables:
- demographics
- distance between user and candidate (zip code centroid)
- habits
- preferences in partner
- how relaxed on preferences
- …
Change in Odds of Male Correspondance Given Preference Relaxation

Change in Odds of Female Correspondance Given Preference Relaxation
cases: hourly energy usage records from >100 buildings over 3 years

variables:
- total energy usage (main load)
- energy usage by component (HVAC, kitchen, lights, fridge, etc.)
- building data: type of building (restaurant, quick serve, retail), state, square feet
- date/time
- …
Model:

- Main load for observation $i$ at building $j$ given by $y_{ij}$, $i \in \{1, \ldots, 36\}$, $j \in \{1, \ldots, 75\}$.

\[
y_{ij} \sim N(\mu_{ij}, \sigma^2) \\
\mu_{ij} = \alpha_{ij} + x_i^T \beta_j \\
\beta_j \sim N(\mu_\beta, \text{diag}(w_1^2, \ldots, w_p^2)) \\
\mu_\beta \sim N(0, \text{diag}(\sigma_{\beta_1}^2, \ldots, \sigma_{\beta_p}^2))
\]

- Non-informative hyper priors on all other parameters
  - $Ga(.01, .01)$ on precisions or $Unif(0, 100)$ on standard deviations
  - $N(0, 1000)$ on mean parameters
Figure: Model performance on a Quick Serve Restaurant (Building 1) and Retail Store (Building 70), showing comparisons between actual values, fitted values, future prediction of Main Load usage for the year 2014, and their 95% CI.