Analysis of longitudinal data to evaluate a policy change

Benjamin French and Patrick J Heagerty

Supplementary Material

In this supplementary material we derive a simple variance estimator for an empirical Bayes estimator. Recall that in an empirical Bayes analysis based on a log-linear model, we obtain an estimate of the average log policy effect by averaging unit-specific log rate ratios:

$$\begin{aligned}\bar{\Delta} &= \frac{1}{n}\sum_{i=1}^{n}\hat{\Delta}_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\log\frac{y_{ij}^A}{\hat{\mu}_{ij}^A}\end{aligned}$$

We wish to obtain a confidence interval for $\exp\bar{\Delta}$, the average policy effect. For simplicity suppose that each unit had one observation collected before and one after the policy change:

$$\begin{aligned}\hat{\Delta}_i &= \log\frac{y_i^A}{\hat{\mu}_i^A} \\ &= \log y_i^A - \log\hat{\mu}_i^A\end{aligned}$$

Recall that the post-policy predicted values $\hat{\mu}_i^A$ are based on a mixed model:

$$\hat{\Delta}_i = \log y_i^A - (\boldsymbol{x}_i^A\hat{\boldsymbol{\beta}}^\star + \boldsymbol{z}_i^A\hat{\boldsymbol{\gamma}}_i + \log N_i^A)$$

To construct a confidence interval for the average log rate ratio, we must estimate its variance:

$$\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\Delta}_i\right] = \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{Var}[\hat{\Delta}_i] + 2\sum_{i\neq i'}^{n}\mathrm{Cov}[\hat{\Delta}_i,\ \hat{\Delta}_{i'}]\right)$$

The covariance terms in this expression are necessary because $\hat{\Delta}_i$ and $\hat{\Delta}_{i'}$ $(i \neq i')$ share common fixed effects, which implies that $\mathrm{Cov}[\hat{\Delta}_i,\ \hat{\Delta}_{i'}] \neq 0$:

$$\begin{aligned}\mathrm{Cov}[\hat{\Delta}_i,\ \hat{\Delta}_{i'}] &= \mathrm{Cov}[\log Y_i^A - (\boldsymbol{x}_i^A\hat{\boldsymbol{\beta}}^\star + \boldsymbol{z}_i^A\hat{\boldsymbol{\gamma}}_i + \log N_i^A), \\ &\qquad \log Y_{i'}^A - (\boldsymbol{x}_{i'}^A\hat{\boldsymbol{\beta}}^\star + \boldsymbol{z}_{i'}^A\hat{\boldsymbol{\gamma}}_{i'} + \log N_{i'}^A)] \\ &= \mathrm{Cov}[\boldsymbol{x}_i^A\hat{\boldsymbol{\beta}}^\star + \boldsymbol{z}_i^A\hat{\boldsymbol{\gamma}}_i,\ \boldsymbol{x}_{i'}^A\hat{\boldsymbol{\beta}}^\star + \boldsymbol{z}_{i'}^A\hat{\boldsymbol{\gamma}}_{i'}] \\ &= \boldsymbol{x}_i^A\boldsymbol{\Sigma}(\boldsymbol{x}_{i'}^A)^T + \boldsymbol{x}_i^A\mathrm{Cov}[\hat{\boldsymbol{\beta}}^\star,\ \hat{\boldsymbol{\gamma}}_{i'}](\boldsymbol{z}_{i'}^A)^T \\ &\quad + \boldsymbol{z}_i^A\mathrm{Cov}[\hat{\boldsymbol{\gamma}}_i,\ \hat{\boldsymbol{\beta}}^\star](\boldsymbol{x}_{i'}^A)^T + \boldsymbol{z}_i^A\boldsymbol{D}(\boldsymbol{z}_{i'}^A)^T\end{aligned}$$

To calculate the covariance between the fixed and random effects, we derive a simple estimator for the random effects. According to Breslow and Clayton [14] $\hat{\boldsymbol{\gamma}}$ is the solution to:

$$\boldsymbol{D}^{-1}\boldsymbol{\gamma} - \sum_{i=1}^{n} \frac{\boldsymbol{z}_i^T(y_i - \mu_i)}{\phi V(\mu_i) g'(\mu_i)} \quad = \quad 0$$

In a log-linear model $g(\mu_i) = \log \mu_i$ and $V(\mu_i) = \mu_i$, so $V(\mu_i)g'(\mu_i) = 1$. If we assume that $\phi = 1$, then the system reduces to:

$$\boldsymbol{D}^{-1}\boldsymbol{\gamma} - \sum_{i=1}^{n} \boldsymbol{z}_i^T(y_i - \mu_i) \quad = \quad 0$$

Therefore $\hat{\boldsymbol{\gamma}}_i = \boldsymbol{D}\boldsymbol{z}_i^T(y_i - \hat{\mu}_i)$. Note that $\boldsymbol{\gamma}_i$ must be iteratively estimated because $\hat{\mu}_i$ depends on $\hat{\boldsymbol{\gamma}}_i$:

$$\hat{\boldsymbol{\gamma}}_i^{(k)} \quad = \quad \boldsymbol{D}\boldsymbol{z}_i^T(y_i - \exp(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}^\star + \boldsymbol{z}_i\hat{\boldsymbol{\gamma}}_i^{(k-1)} + \log N_i))$$

We use a one-step estimator for $\boldsymbol{\gamma}_i$ and assume that $\hat{\boldsymbol{\gamma}}_i^{(0)} = 0$:

$$\hat{\boldsymbol{\gamma}}_i^{(1)} \quad = \quad \boldsymbol{D}\boldsymbol{z}_i^T(y_i - \exp(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}^\star + \log N_i))$$

Recall that $\boldsymbol{\gamma}_i$ is estimated using the data observed before the policy change:

$$\hat{\boldsymbol{\gamma}}_i^{(1)} \quad = \quad \boldsymbol{D}(\boldsymbol{z}_i^B)^T(y_i^B - \exp(\boldsymbol{x}_i^B\hat{\boldsymbol{\beta}}^\star + \log N_i^B))$$

We use a Taylor series expansion of $\exp(\boldsymbol{x}_i^B\hat{\boldsymbol{\beta}}^\star)$ to obtain an estimator for $\boldsymbol{\gamma}_i$ that is linear in $\hat{\boldsymbol{\beta}}^\star$:

$$\hat{\boldsymbol{\gamma}}_i^{(1)} \quad \approx \quad \boldsymbol{D}(\boldsymbol{z}_i^B)^T(y_i^B - \exp(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \log N_i^B) - \exp(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \log N_i^B)\boldsymbol{x}_i^B(\hat{\boldsymbol{\beta}}^\star - \boldsymbol{\beta}^\star))$$

With the linearized one-step estimator we return to our earlier covariance calculations:

$$\begin{aligned}
\text{Cov}[\hat{\boldsymbol{\beta}}^\star,\ \hat{\boldsymbol{\gamma}}_{i'}] \quad &\approx \quad \text{Cov}[\hat{\boldsymbol{\beta}}^\star,\ \hat{\boldsymbol{\gamma}}_{i'}^{(1)}] \\
&= \quad -\text{Cov}[\hat{\boldsymbol{\beta}}^\star,\ \exp(\boldsymbol{x}_{i'}^B\boldsymbol{\beta}^\star + \log N_{i'}^B)\boldsymbol{x}_{i'}^B\hat{\boldsymbol{\beta}}^\star]\boldsymbol{z}_{i'}^B\boldsymbol{D} \\
&= \quad -\exp(\boldsymbol{x}_{i'}^B\boldsymbol{\beta}^\star + \log N_{i'}^B)\boldsymbol{\Sigma}(\boldsymbol{x}_{i'}^B)^T\boldsymbol{z}_{i'}^B\boldsymbol{D}
\end{aligned}$$

$$\begin{aligned}
\text{Cov}[\hat{\boldsymbol{\gamma}}_i,\ \hat{\boldsymbol{\beta}}^\star] \quad &\approx \quad \text{Cov}[\hat{\boldsymbol{\gamma}}_i^{(1)},\ \hat{\boldsymbol{\beta}}^\star] \\
&= \quad -\boldsymbol{D}(\boldsymbol{z}_i^B)^T\text{Cov}[\exp(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \log N_i^B)\boldsymbol{x}_i^B\hat{\boldsymbol{\beta}}^\star,\ \hat{\boldsymbol{\beta}}^\star] \\
&= \quad -\exp(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \log N_i^B)\boldsymbol{D}(\boldsymbol{z}_i^B)^T\boldsymbol{x}_i^B\boldsymbol{\Sigma}
\end{aligned}$$

Therefore the covariance between two unit-specific log rate ratios is:

$$\begin{aligned}
\text{Cov}[\hat{\Delta}_i,\ \hat{\Delta}_{i'}] \quad \approx \quad & \boldsymbol{x}_i^A\boldsymbol{\Sigma}(\boldsymbol{x}_{i'}^A)^T - \exp(\boldsymbol{x}_{i'}^B\boldsymbol{\beta}^\star + \log N_{i'}^B)\boldsymbol{x}_i^A\boldsymbol{\Sigma}(\boldsymbol{x}_{i'}^B)^T\boldsymbol{z}_{i'}^B\boldsymbol{D}(\boldsymbol{z}_{i'}^A)^T \\
& -\ \exp(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \log N_i^B)\boldsymbol{z}_i^A\boldsymbol{D}(\boldsymbol{z}_i^B)^T\boldsymbol{x}_i^B\boldsymbol{\Sigma}(\boldsymbol{x}_{i'}^A)^T + \boldsymbol{z}_i^A\boldsymbol{D}(\boldsymbol{z}_{i'}^A)^T
\end{aligned}$$

Estimation requires consistent estimates of $\boldsymbol{\beta}^\star$, $\boldsymbol{\Sigma}$, and $\boldsymbol{D}$, which are typically available from standard mixed effects regression output.

To construct a confidence interval for the average log rate ratio, we must also calculate the variance of each unit-specific log rate ratio:

$$\begin{aligned}
\text{Var}[\hat{\Delta}_i] &= \text{Var}[\log Y_i^A - (\boldsymbol{x}_i^A\hat{\boldsymbol{\beta}}^\star + \boldsymbol{z}_i^A\hat{\boldsymbol{\gamma}}_i + \log N_i^A)] \\
&= \text{Var}[\log Y_i^A] + \text{Var}[\boldsymbol{x}_i^A\hat{\boldsymbol{\beta}}^\star + \boldsymbol{z}_i^A\hat{\boldsymbol{\gamma}}_i + \log N_i^A] \\
&\quad - 2\text{Cov}[\log Y_i^A,\ \boldsymbol{x}_i^A\hat{\boldsymbol{\beta}}^\star + \boldsymbol{z}_i^A\hat{\boldsymbol{\gamma}}_i + \log N_i^A] \\
&= \text{Var}[\log Y_i^A] + \text{Var}[\boldsymbol{x}_i^A\hat{\boldsymbol{\beta}}^\star] + \text{Var}[\boldsymbol{z}_i^A\hat{\boldsymbol{\gamma}}_i] \\
&\quad + 2\text{Cov}[\boldsymbol{x}_i^A\hat{\boldsymbol{\beta}}^\star,\ \boldsymbol{z}_i^A\hat{\boldsymbol{\gamma}}_i] - 2\text{Cov}[\log Y_i^A,\ \boldsymbol{x}_i^A\hat{\boldsymbol{\beta}}^\star + \boldsymbol{z}_i^A\hat{\boldsymbol{\gamma}}_i + \log N_i^A] \\
&= \text{Var}[\log Y_i^A] + \boldsymbol{x}_i^A\boldsymbol{\Sigma}(\boldsymbol{x}_i^A)^T + \boldsymbol{z}_i^A\boldsymbol{D}(\boldsymbol{z}_i^A)^T \\
&\quad + 2\boldsymbol{x}_i^A\text{Cov}[\hat{\boldsymbol{\beta}}^\star,\ \hat{\boldsymbol{\gamma}}_i](\boldsymbol{z}_i^A)^T - 2\text{Cov}[\log Y_i^A,\ \boldsymbol{x}_i^A\hat{\boldsymbol{\beta}}^\star + \boldsymbol{z}_i^A\hat{\boldsymbol{\gamma}}_i + \log N_i^A]
\end{aligned}$$

We use a Taylor series expansion of $\log y_i^A$ to calculate $\text{Var}(\log Y_i^A)$:

$$\begin{aligned}
\text{Var}[\log Y_i^A] &\approx \text{Var}[\log \text{E}[Y_i^A] + (Y_i^A - \text{E}[Y_i^A])/\text{E}[Y_i^A]] \\
&= \text{Var}[Y_i^A]/(\text{E}[Y_i^A])^2 \\
&= \exp(-2(\boldsymbol{x}_i^A\boldsymbol{\beta}^\star + \boldsymbol{z}_i^A\boldsymbol{D}(\boldsymbol{z}_i^A)^T/2 + \log N_i^A))\text{Var}[Y_i^A]
\end{aligned}$$

We do not directly model $\text{Var}[Y_i^A]$, but we estimate it based on a model for $\text{Var}[Y_i^B]$. Recall the hierarchical model for the data observed before the policy change:

$$\begin{aligned}
Y_i^B \mid \boldsymbol{\gamma}_i &\sim P(\exp(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \boldsymbol{z}_i^B\boldsymbol{\gamma}_i + \log N_i^B)) \\
\boldsymbol{\gamma}_i &\sim N_2(0, \boldsymbol{D})
\end{aligned}$$

According to the conditional variance formula:

$$\begin{aligned}
\text{Var}[Y_i^B] &= \text{E}_\gamma[\text{Var}_Y[Y_i^B \mid \boldsymbol{\gamma}_i]] + \text{Var}_\gamma[\text{E}_Y[Y_i^B \mid \boldsymbol{\gamma}_i]] \\
&= \text{E}_\gamma[\exp(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \boldsymbol{z}_i^B\boldsymbol{\gamma}_i + \log N_i^B)] \\
&\quad + \text{Var}_\gamma[\exp(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \boldsymbol{z}_i^B\boldsymbol{\gamma}_i + \log N_i^B)] \\
&= \exp(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \log N_i^B)\text{E}_\gamma[\exp(\boldsymbol{z}_i^B\boldsymbol{\gamma}_i)] \\
&\quad + \exp(2(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \log N_i^B))\text{Var}_\gamma[\exp(\boldsymbol{z}_i^B\boldsymbol{\gamma}_i)]
\end{aligned}$$

We use the moment generating function for a Normal random variable:

$$\begin{aligned}
\text{E}_\gamma[\exp(\boldsymbol{z}_i^B\boldsymbol{\gamma}_i)] &= \exp(\boldsymbol{z}_i^B\text{E}_\gamma[\boldsymbol{\gamma}_i] + \boldsymbol{z}_i^B\text{Var}_\gamma[\gamma_i](\boldsymbol{z}_i^B)^T/2) \\
&= \exp(\boldsymbol{z}_i^B\boldsymbol{D}(\boldsymbol{z}_i^B)^T/2)
\end{aligned}$$

$$\begin{aligned}
\text{Var}_\gamma[\exp(\boldsymbol{z}_i^B\boldsymbol{\gamma}_i)] &= \text{E}_\gamma[(\exp(\boldsymbol{z}_i^B\boldsymbol{\gamma}_i))^2] - (\text{E}_\gamma[\exp(\boldsymbol{z}_i^B\boldsymbol{\gamma}_i)])^2 \\
&= \exp(2\boldsymbol{z}_i^B\text{E}_\gamma[\boldsymbol{\gamma}_i] + 2\boldsymbol{z}_i^B\text{Var}_\gamma[\boldsymbol{\gamma}_i](\boldsymbol{z}_i^B)^T) - (\exp(\boldsymbol{z}_i^B\boldsymbol{D}(\boldsymbol{z}_i^B)^T/2))^2 \\
&= \exp(\boldsymbol{z}_i^B\boldsymbol{D}(\boldsymbol{z}_i^B)^T)(\exp(\boldsymbol{z}_i^B\boldsymbol{D}(\boldsymbol{z}_i^B)^T) - 1)
\end{aligned}$$

With these expressions we return to our earlier variance calculation:

$$\begin{aligned}
\mathrm{Var}[Y_i^B] &= \exp(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \boldsymbol{z}_i^B\boldsymbol{D}(\boldsymbol{z}_i^B)^T/2 + \log N_i^B) \\
&\quad + \exp(2(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \boldsymbol{z}_i^B\boldsymbol{D}(\boldsymbol{z}_i^B)^T/2 + \log N_i^B))(\exp(\boldsymbol{z}_i^B\boldsymbol{D}(\boldsymbol{z}_i^B)^T) - 1) \\
&= \exp(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \boldsymbol{z}_i^B\boldsymbol{D}(\boldsymbol{z}_i^B)^T/2 + \log N_i^B) \\
&\quad \times [1 + (\exp(\boldsymbol{z}_i^B\boldsymbol{D}(\boldsymbol{z}_i^B)^T) - 1)\exp(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \boldsymbol{z}_i^B\boldsymbol{D}(\boldsymbol{z}_i^B)^T/2 + \log N_i^B)] \\
&\equiv \mathrm{E}[Y_i^B](1 + \phi\mathrm{E}[Y_i^B])
\end{aligned}$$

We recognize this as the standard variance for an over-dispersed Poisson random variable, where the dispersion parameter $\phi = \exp(\boldsymbol{z}_i^B\boldsymbol{D}(\boldsymbol{z}_i^B)^T) - 1$. Based on this model for $\mathrm{Var}[Y_i^B]$ we obtain:

$$\begin{aligned}
\mathrm{Var}[\log Y_i^A] &\approx \exp(-(\boldsymbol{x}_i^A\boldsymbol{\beta}^\star + \boldsymbol{z}_i^A\boldsymbol{D}(\boldsymbol{z}_i^A)^T/2 + \log N_i^A)) \\
&\quad \times [1 + (\exp(\boldsymbol{z}_i^A\boldsymbol{D}(\boldsymbol{z}_i^A)^T) - 1)\exp(\boldsymbol{x}_i^A\boldsymbol{\beta}^\star + \boldsymbol{z}_i^A\boldsymbol{D}(\boldsymbol{z}_i^A)^T/2 + \log N_i^A)] \\
&\equiv (1 + \phi\mathrm{E}[Y_i^A])/\mathrm{E}[Y_i^A]
\end{aligned}$$

It remains to calculate the covariance between $\log Y_i^A$ and $\hat{\mu}_i^A$:

$$\begin{aligned}
\mathrm{Cov}[\log Y_i^A,\ \hat{\mu}_i^A] &= \mathrm{E}_\gamma[\mathrm{Cov}[\log Y_i^A,\ \boldsymbol{x}_i^A\hat{\boldsymbol{\beta}}^\star + \boldsymbol{z}_i^A\hat{\boldsymbol{\gamma}}_i + \log N_i^A \mid \boldsymbol{\beta}^\star, \boldsymbol{\gamma}_i]] \\
&\quad + \mathrm{Cov}_\gamma[\mathrm{E}[\log Y_i^A \mid \boldsymbol{\beta}^\star, \boldsymbol{\gamma}_i],\ \mathrm{E}[\boldsymbol{x}_i^A\hat{\boldsymbol{\beta}}^\star + \boldsymbol{z}_i^A\hat{\boldsymbol{\gamma}}_i + \log N_i^A \mid \boldsymbol{\beta}^\star, \boldsymbol{\gamma}_i]] \\
&\approx 0 + \mathrm{Cov}_\gamma[\boldsymbol{x}_i^A\boldsymbol{\beta}^\star + \boldsymbol{z}_i^A\boldsymbol{\gamma}_i + \log N_i^A,\ \boldsymbol{x}_i^A\boldsymbol{\beta}^\star + \boldsymbol{z}_i^A\boldsymbol{\gamma}_i + \log N_i^A] \\
&= \boldsymbol{z}_i^A\boldsymbol{D}(\boldsymbol{z}_i^A)^T
\end{aligned}$$

Therefore the variance of each unit-specific log rate ratio is:

$$\begin{aligned}
\mathrm{Var}[\hat{\Delta}_i] &\approx \mathrm{Var}[\log Y_i^A] + \boldsymbol{x}_i^A\boldsymbol{\Sigma}(\boldsymbol{x}_i^A)^T - \boldsymbol{z}_i^A\boldsymbol{D}(\boldsymbol{z}_i^A)^T \\
&\quad - 2\exp(\boldsymbol{x}_i^B\boldsymbol{\beta}^\star + \log N_i^B)\boldsymbol{x}_i^A\boldsymbol{\Sigma}(\boldsymbol{x}_i^B)^T\boldsymbol{z}_i^B\boldsymbol{D}(\boldsymbol{z}_i^A)^T
\end{aligned}$$

Estimation requires consistent estimates of $\boldsymbol{\beta}^\star$, $\boldsymbol{\Sigma}$, and $\boldsymbol{D}$.

Hence a $(1-\alpha)\%$ confidence interval for the average policy effect is:

$$\exp\left(\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Delta}_i\right) \pm t_{\alpha/2,n-1}\sqrt{\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\Delta}_i\right]}\right)$$

To derive this confidence interval we assumed that each study unit had one observation collected before and one after the policy change. In practice multiple observations are collected before and after the policy change, as in our case study. In this case a simple approach is to substitute $\bar{\boldsymbol{x}}_i^B$, $\bar{\boldsymbol{x}}_i^A$, $\bar{\boldsymbol{z}}_i^B$, $\bar{\boldsymbol{z}}_i^A$, $\bar{N}_i^B$, and $\bar{N}_i^A$ for $\boldsymbol{x}_i^B$, $\boldsymbol{x}_i^A$, $\boldsymbol{z}_i^B$, $\boldsymbol{z}_i^A$, N_i^B, and N_i^A, respectively.