Module II: An Example of a Two-stage Model; NMMAPS Study

Francesca Dominici

and

Scott L. Zeger

NMMAPS Example of Two-Stage Hierarchical Model

- National Morbidity and Mortality Air Pollution Study (NMMAPS)
- Daily data on cardiovascular/respiratory mortality in 10 largest cities in U.S.
- Daily particulate matter (PM10) data
- Log-linear regression estimate relative risk of mortality per 10 unit increase in PM10 for each city
- Estimate and statistical standard error for each city

Semi-Parametric Poisson Regression

$$\log \mu_t = \beta_0 + \beta x_{t-\ell} + s_1(\text{time}, \lambda_1) + s_2(\text{temp}, \lambda_2) + \text{others}$$

Log-Relative Rate

Season

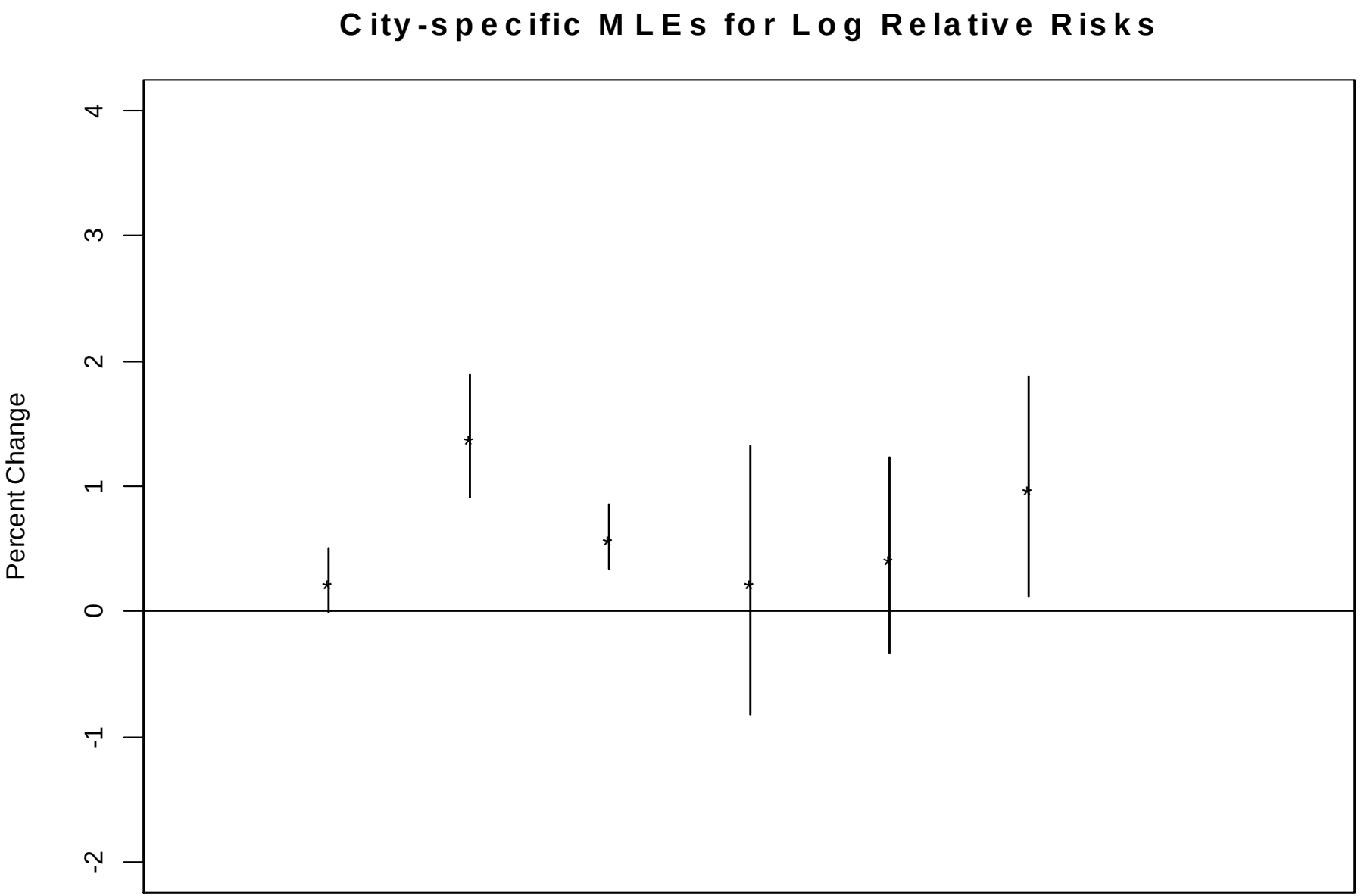
Weather

Splines

Relative Risks* for Six Largest Cities

City	RR Estimate (% per 10 micrograms/ml	Statistical Standard Error	Statistical Variance
Los Angeles	0.25	0.13	.0169
New York	1.4	0.25	.0625
Chicago	0.60	0.13	.0169
Dallas/Ft Worth	0.25	0.55	.3025
Houston	0.45	0.40	.1600
San Diego	1.0	0.45	.2025

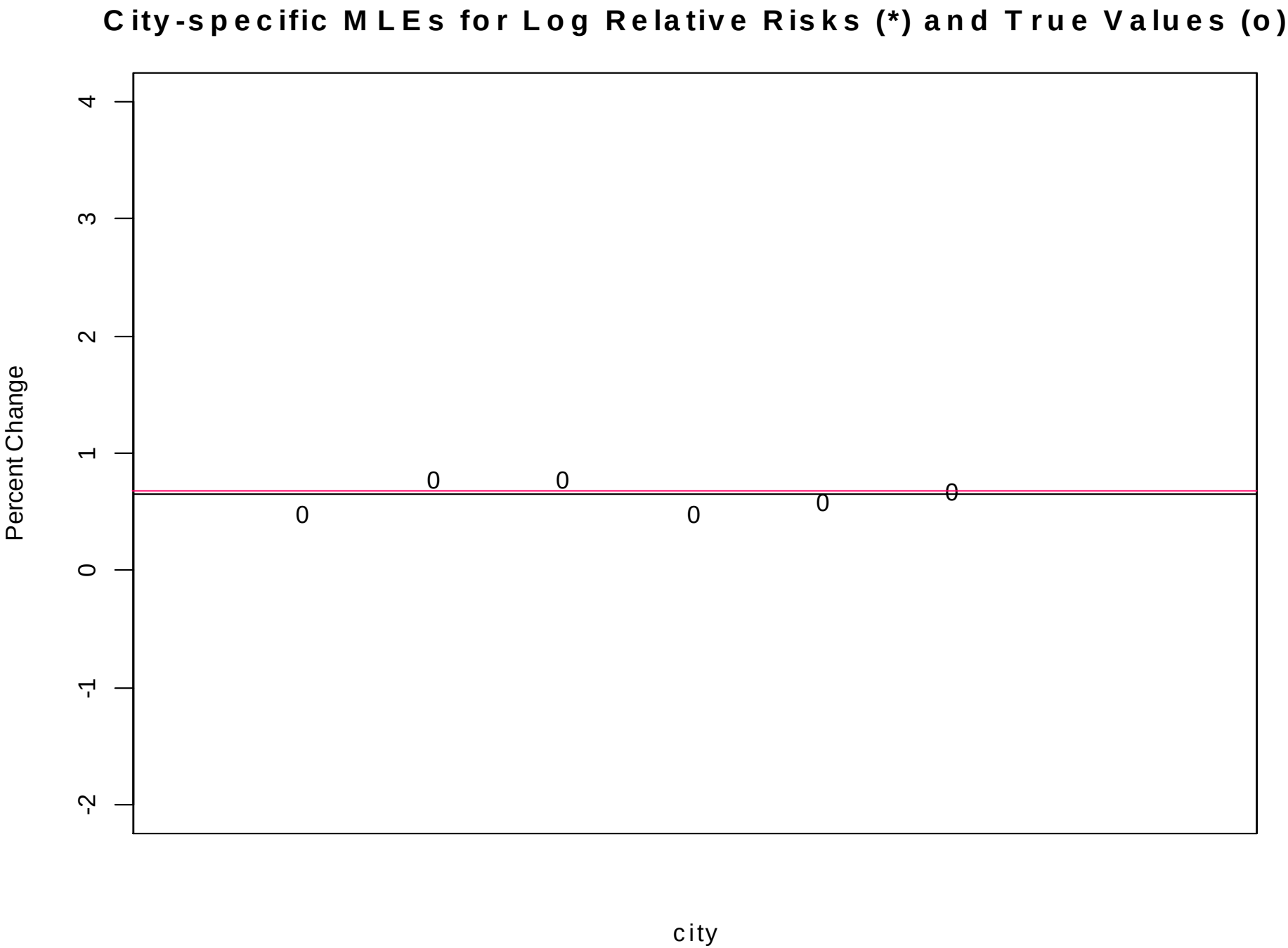
Approximate values read from graph in Daniels, et al. 2000. AJE

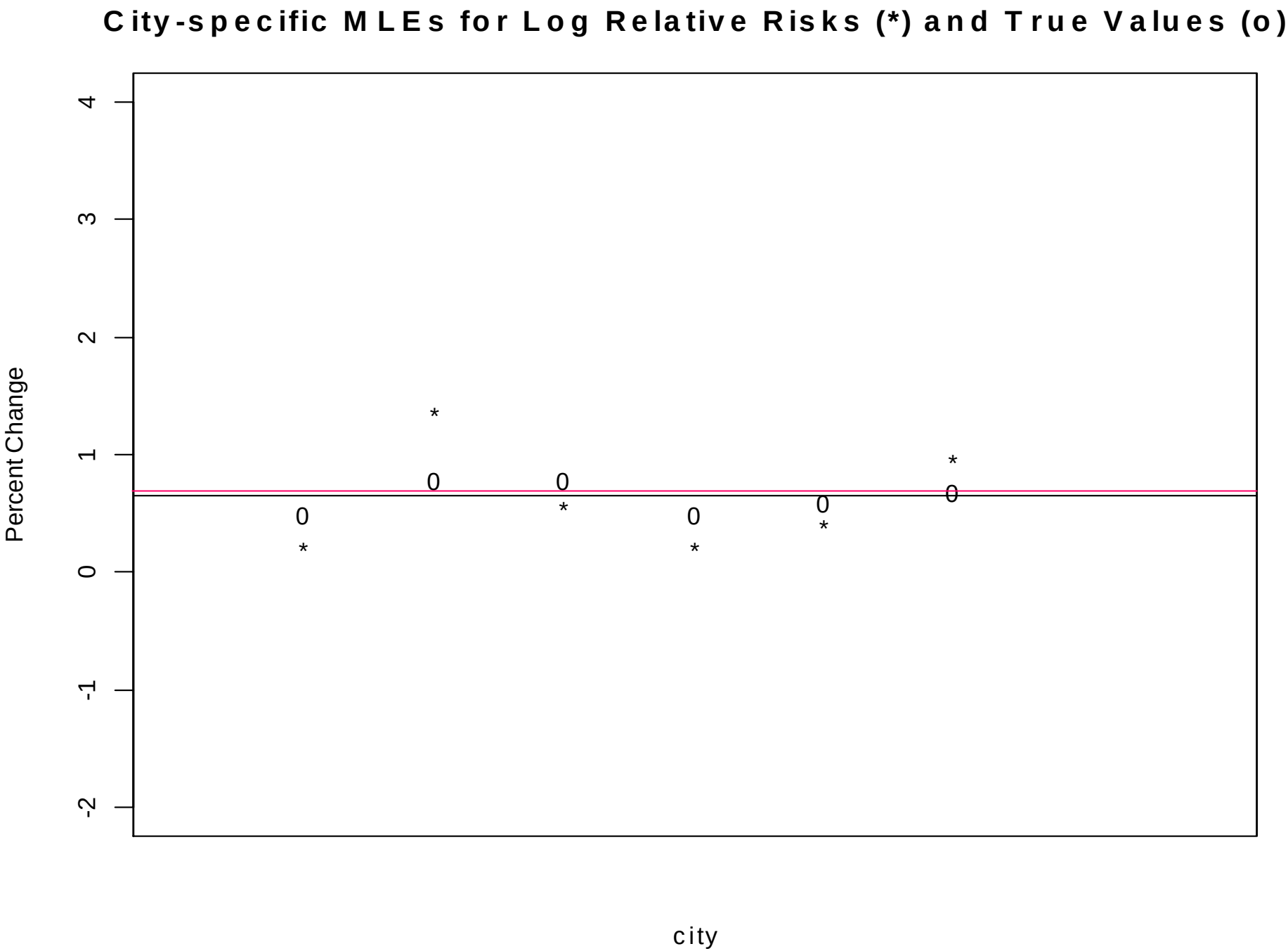

City-specific MLEs for Log Relative Risks
Percent Change
4
3
2
1
0
-1
-2

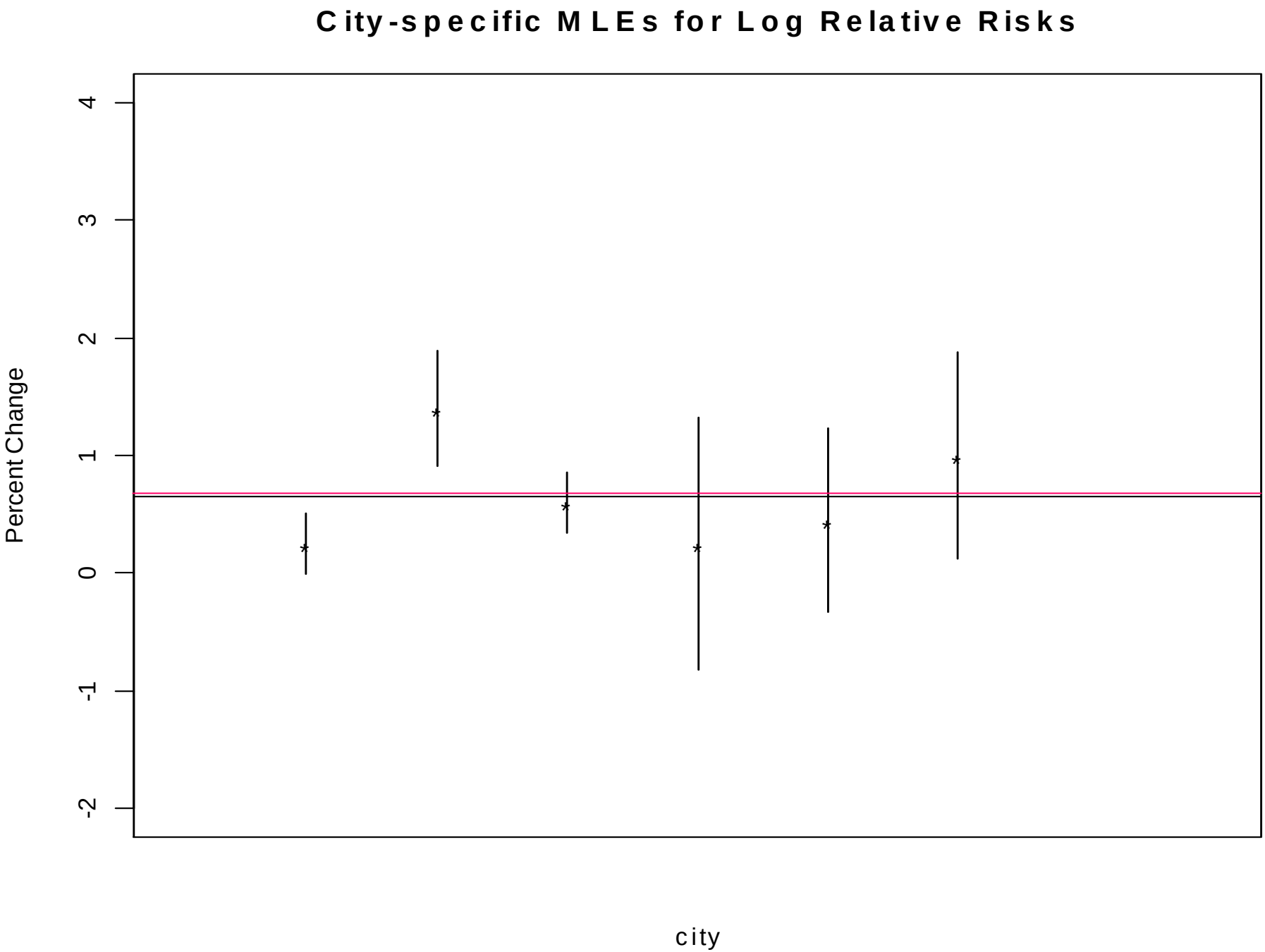
- $\hat{\beta}_c$ is the city-specific relative risk estimate
- β_c is the "true" city-specific relative risk
- α is the average relative risk over all the cities
- e_c is the statistical error
- d_c is the deviation of the true city-specific relative risk from the overall mean

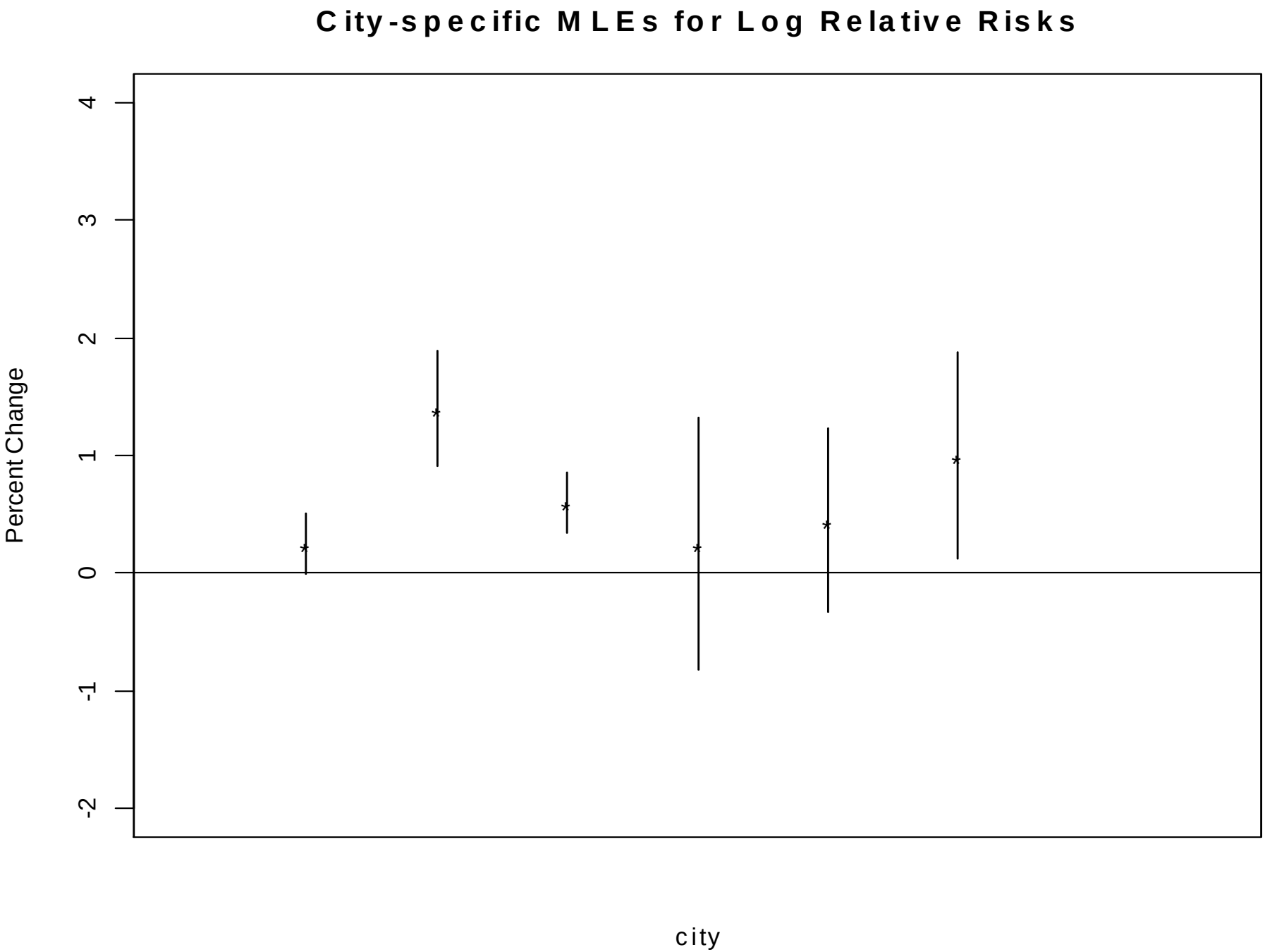
Sources of Variation

- $\hat{\beta}_c = \beta_c + \epsilon_c$
- $\beta_c = \alpha + d_c$
- $\hat{\beta}_c = \alpha + d_c + \epsilon_c$
- $var(\hat{\beta}_c) = var(\epsilon_c) + var(d_c) = v_c + NV$


City-specific MLEs for Log Relative Risks (*) and True Values (o)
Percent Change
4
3
2
1
0
-1
-2
0
0
0
0
0
0
city


City-specific MLEs for Log Relative Risks (*) and True Values (o)
Percent Change
4
3
2
1
0
-1
-2
city


City-specific MLEs for Log Relative Risks
Percent Change
4
3
2
1
0
-1
-2
city


City-specific MLEs for Log Relative Risks
Percent Change
4
3
2
1
0
-1
-2
city

Notation

- v_c is the statistical variance
- e_c is the statistical error
- d_c is the deviation of city-specific true relative risk from the average
- $var(e_c)$ is the random noise
- $NV = var(d_c)$ is the variance of the true relative risks across cities, also called natural variance or heterogeneity
- $TV_c = NV + var(e_c)$ is the total variance

Estimating Overall Mean

- Idea: give more weight to more precise values
- Specifically, weight estimates inversely proportional to their variances

Estimating the Overall Mean

We can estimate the average relative risk over all the cities by:

$$\begin{aligned}\hat{\alpha} &= \left(\sum_c w_c\right)^{-1} \sum_c w_c \hat{\beta}_c \\ \mathrm{var}(\hat{\alpha}) &= \left(\sum_c w_c\right)^{-1}\end{aligned}$$

The weights are calculated as follows:

$$\begin{aligned}h_c &= 1/(v_c + \widehat{NV}) \\ w_c &= h_c / \sum_c h_c, \ \sum_c w_c = 1 \\ \widehat{NV} &= \frac{1}{N-1} \sum_c (\hat{\beta}_c - \bar{\beta})^2 \\ &= \text{variance across cities of } \hat{\beta}_c - \frac{1}{N} \sum_c v_c\end{aligned}$$

Calculations for Empirical Bayes Estimates

City	Log RR (bc)	Stat Var (vc)	Total Var (TVc)	1/TVc	wc
LA	0.25	.0169	.0994	10.1	.27
NYC	1.4	.0625	.145	6.9	.18
Chi	0.60	.0169	.0994	10.1	.27
Dal	0.25	.3025	.385	2.6	.07
Hou	0.45	.160	,243	4.1	.11
SD	1.0	.2025	.285	3.5	.09
Over-all	0.65			37.3	1.00

$\alpha = .27 * 0.25 + .18 * 1.4 + .27 * 0.60 + .07 * 0.25 + .11 * 0.45 + 0.9 * 1.0 = 0.65$

$Var(\alpha) = 1/Sum(1/TVc) = 0.164^2$

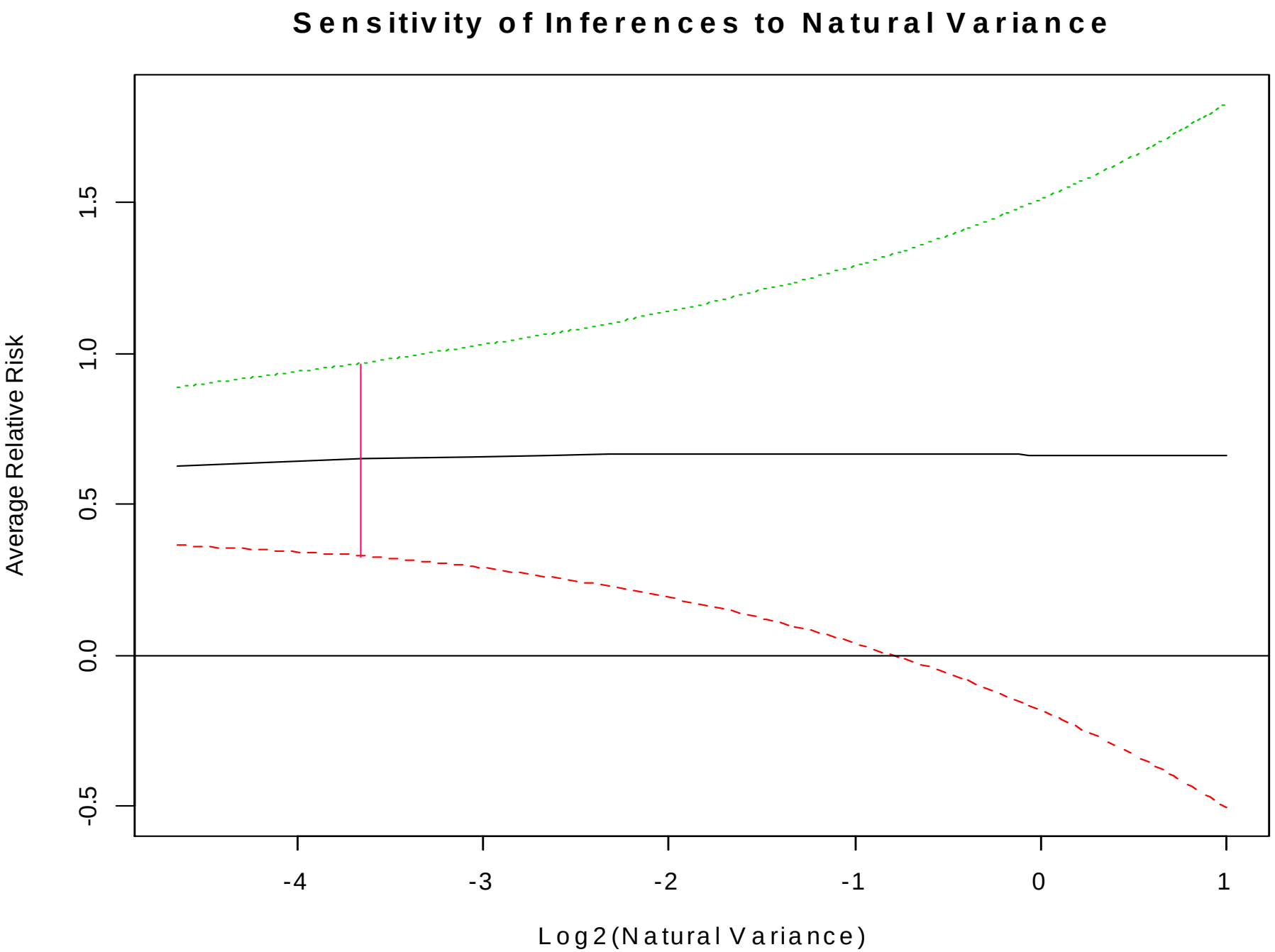
Software in R

```
beta.hat <-c(0.25,1.4,0.50,0.25,0.45,1.0)
se <- c(0.13,0.25,0.13,0.55,0.40,0.45)
NV <- var(beta.hat) - mean(se^2)
TV <- se^2 + NV
tmp<- 1/TV
ww <- tmp/sum(tmp)
v.alphahat <- sum(ww)^{-1}
alpha.hat <- v.alphahat*sum(beta.hat*ww)
```

Two Extremes

- Natural variance >> Statistical variances
  - Weights wc approximately constant = 1/n
  - Use ordinary mean of estimates regardless of their relative precision

- Statistical variances >> Natural variance
  - Weight each estimator inversely proportional to its statistical variance


Sensitivity of Inferences to Natural Variance
Average Relative Risk
1.5
1.0
0.5
0.0
-0.5
-4
-3
-2
-1
0
1
Log2(Natural Variance)

Estimating Relative Risk for Each City

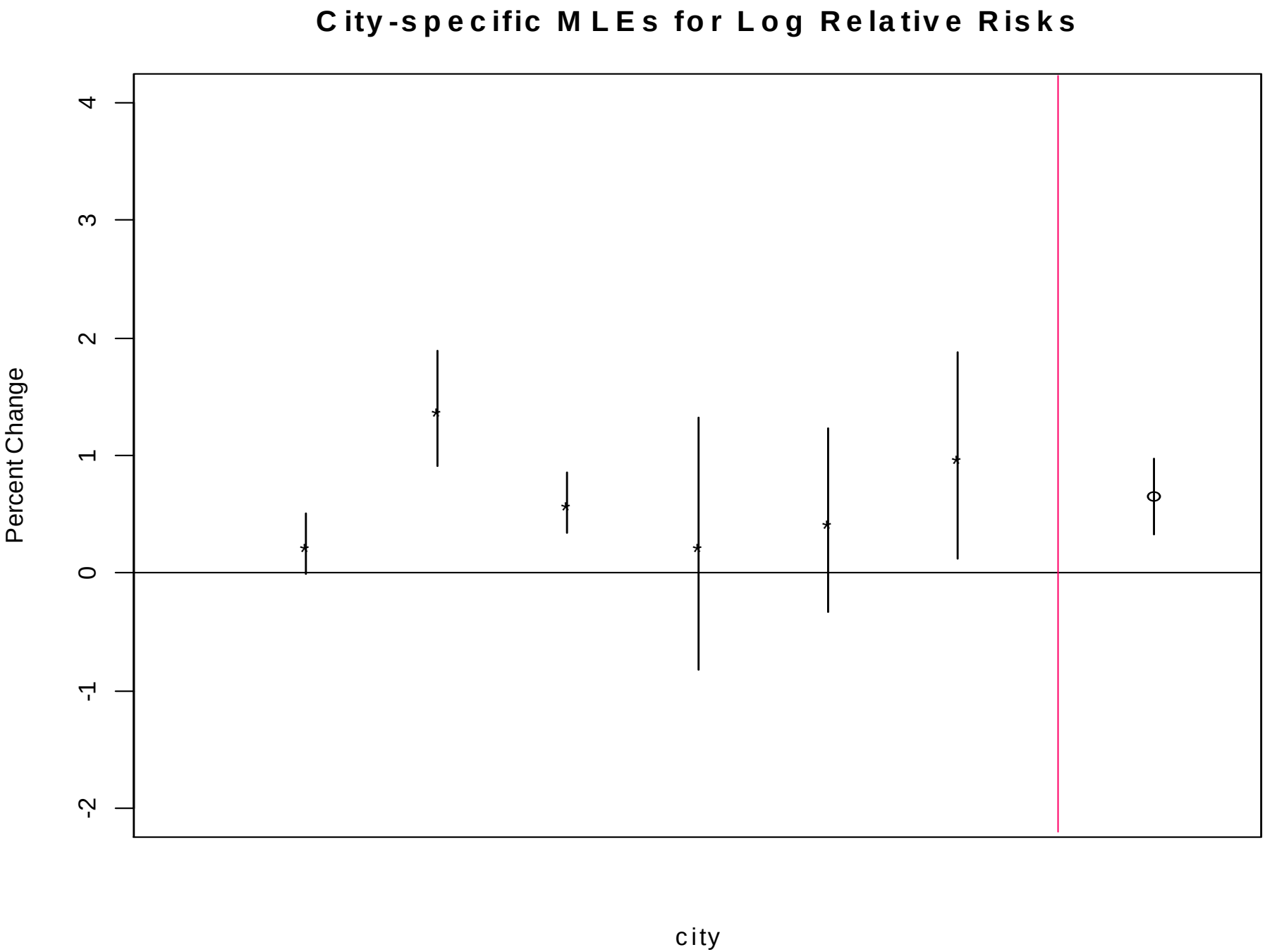
- Disease screening analogy
  - Test result from imperfect test
  - Positive predictive value combines prevalence with test result using Bayes theorem
- Empirical Bayes estimator of the true value for a city is the conditional expectation of the true value given the data $\boxed{E(\beta_c \mid \hat{\beta}_c)}$

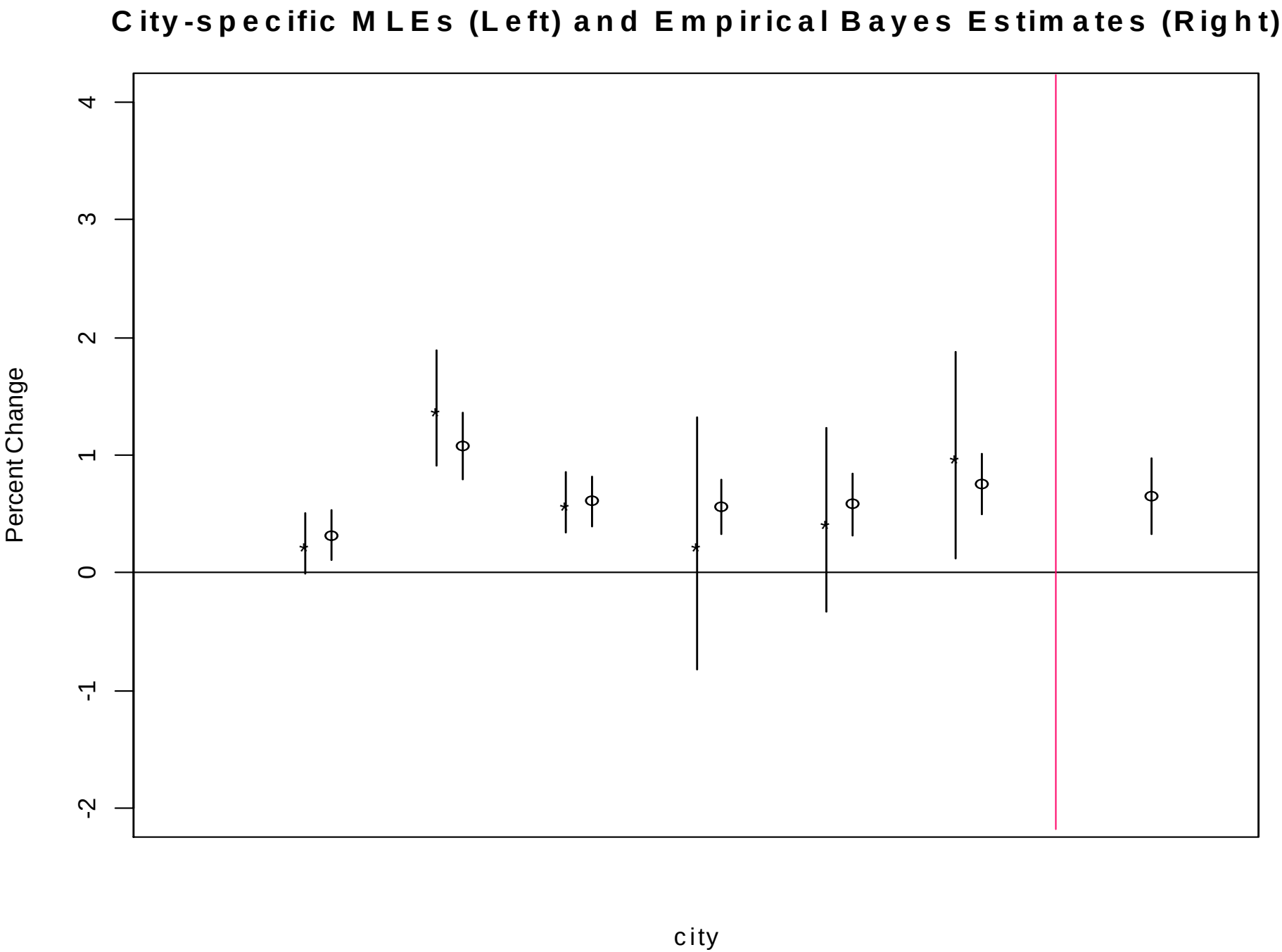
Empirical Bayes Estimate

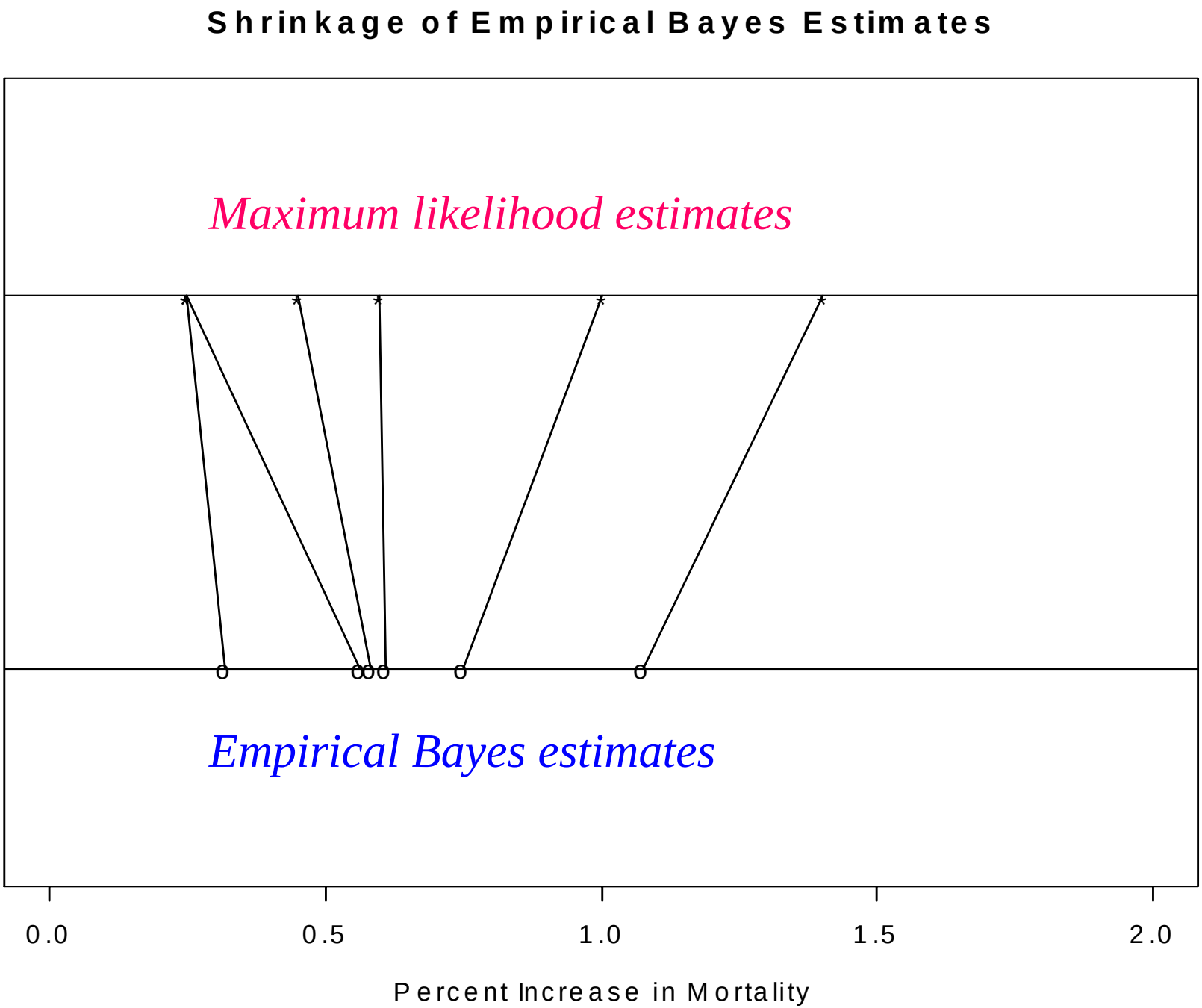
- $\tilde{\beta}_c = E[\beta_c \mid \hat{\beta}_c] = \theta_c \hat{\beta}_c + (1 - \theta_c)\hat{\alpha}$
- $\theta_c = NV/(NV + v_c)$
- $var(\tilde{\beta}_c) \simeq \theta_c v_c$

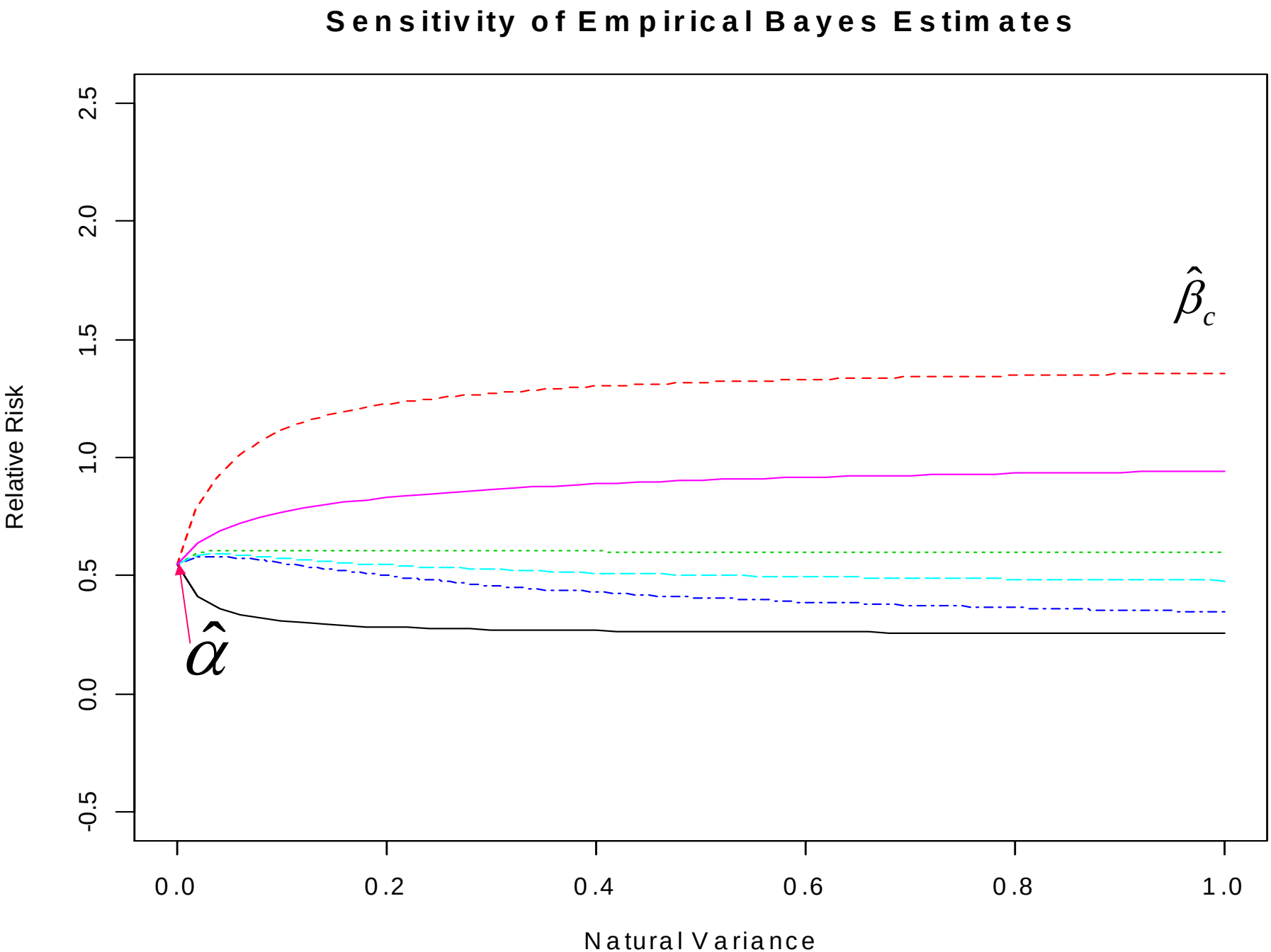
Calculations for Empirical Bayes Estimates

City	Log RR $\hat{\beta}_c$	Stat Var (vc)	Total Var (TVc)	1/TVc	wc	θ_c	RR.EB $\tilde{\beta}_c$	se RR.EB
LA	0.25	.0169	.0994	10.1	.27	.83	0.32	0.17
NYC	1.4	.0625	.145	6.9	.18	.57	1.1	0.14
Chi	0.60	.0169	.0994	10.1	.27	.83	0.61	0.11
Dal	0.25	.3025	.385	2.6	.07	.21	0.56	0.12
Hou	0.45	.160	,243	4.1	.11	.34	0.58	0.14
SD	1.0	.2025	.285	3.5	.09	.29	0.75	0.13
Over-all	0.65	1/37.3= 0.027		37.3	1.00		0.65	0.16


City-specific MLEs for Log Relative Risks
Percent Change
4
3
2
1
0
-1
-2
city


City-specific MLEs (Left) and Empirical Bayes Estimates (Right)
Percent Change
4
3
2
1
0
-1
-2
city


Shrinkage of Empirical Bayes Estimates
Maximum likelihood estimates
Empirical Bayes estimates
0.0
0.5
1.0
1.5
2.0
Percent Increase in Mortality


Sensitivity of Empirical Bayes Estimates
Relative Risk
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
0.0
0.2
0.4
0.6
0.8
1.0
Natural Variance
$\hat{\beta}_c$
$\hat{\alpha}$

Key Ideas

- Better to use data for all cities to estimate the relative risk for a particular city
  - Reduce variance by adding some bias
  - Smooth compromise between city specific estimates and overall mean
- Empirical-Bayes estimates depend on measure of natural variation
  - Assess sensivity to estimate of NV

Caveats

- Used simplistic methods to illustrate the key ideas:
  - Treated natural variance and overall estimate as known when calculating uncertainty in EB estimates
  - Assumed normal distribution or true relative risks
- Can do better using Markov Chain Monte Carlo methods – more to come