Lab 1: NMMAPS

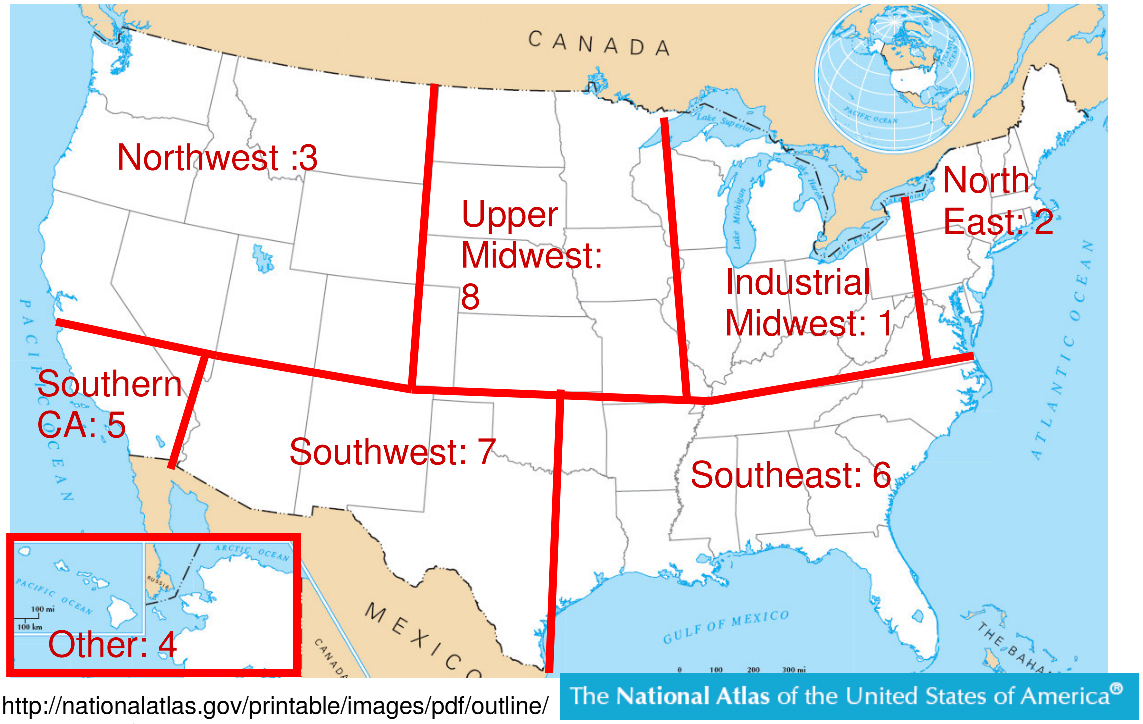
Data: City-specific air pollution effect estimates from an analysis of a 100 city subset of the **N**ational **M**orbidity **M**ortality and **A**ir **P**ollution **S**tudy (NMMAPS) and city-level covariates. NMMAPS is a multi-city study containing city-specific daily time-series data on air pollution levels (including particulate matter of 10 microns and less in aerodynamic diameter – PM_{10}), mortality counts, temperature and dewpoint temperature.

Variables

- city: four letter city name abbreviation
- cityname: human readable name of the city
- state: state in which the city is located
- beta: city-specific coefficient on lag 1 PM_{10} from a log-linear model relating changes in daily particulate matter air pollution to changes in daily mortality
- sebeta: standard error of beta
- population2000: city's population from 2000 census
- pdrive: proportion of the population that drives to work
- punem: proportion of the population unemployed
- pdeg: proportion of the population with a college degree or higher
- p65p: proportion of the city's population aged 65 or older
- latitude
- longitude
- altitude (contains some missing values)
- region: one of 7 NMMAPS regions (Industrial Midwest = IM, North East = NE, North West = NW, Southern California = SC, South East = SE, South West = SW, Upper Midwest = UM)

Goal: Estimate regional average associations between daily variations in PM_{10} and daily variations in city-level mortality counts by combining city-specific estimates of log relative risks using shrinkage.

For now, we will ignore the standard error of the betas and consider only the variables beta and region.

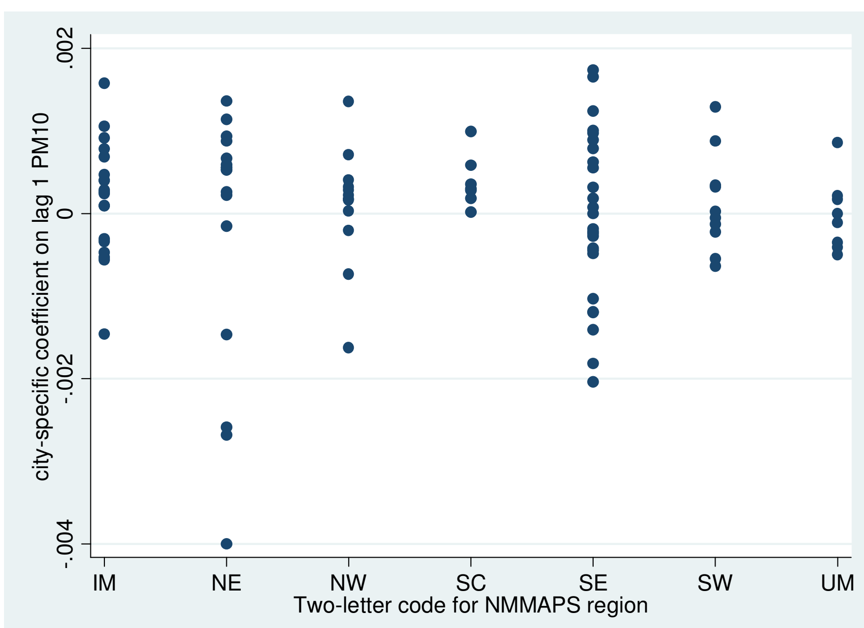


http://nationalatlas.gov/printable/images/pdf/outline/states(u).pdf

Exploratory Data Analysis

```
. sort region
```

```
. scatter beta region, xlabel(1 2 3 4 5 6 7, valuelabel)
```



Approach A: No shrinkage
Calculate each region's observed average coefficient on PM

$\overline{\beta}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\beta_{ij}$ where β_{ij} is the city-specific estimate and *j* indexes region while *i* indexes city within region and n_j is the number of cities in region *j*.

```
. mean beta, over(region)

Mean estimation                         Number of obs    =     100

             IM: region = IM
             NE: region = NE
             NW: region = NW
             SC: region = SC
             SE: region = SE
             SW: region = SW
             UM: region = UM

--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
beta         |
          IM |   .0001913   .0001602     -.0001265    .0005092
          NE |   -.000156   .0003765     -.0009031    .0005912
          NW |   .0001014   .0001951     -.0002858    .0004886
          SC |   .0003898   .0001203      .0001512    .0006285
          SE |  -.0000337   .0001989     -.0004283    .0003609
          SW |    .000129   .0001911     -.0002501    .0005081
          UM |  -.0000138   .0001564     -.0003241    .0002965
--------------------------------------------------------------
```

Approach B: Complete shrinkage
Calculate overall average of city-specific estimates

$$\overline{\beta} = \frac{\sum_{j=1}^{J}\frac{n_j}{\sigma^2}\overline{\beta}_j}{\sum_{j=1}^{J}\frac{n_j}{\sigma^2}} = \frac{\sum_{j=1}^{J}n_j\overline{\beta}_j}{\sum_{j=1}^{J}n_j} = \frac{\sum_{j=1}^{J}n_j\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\beta_{ij}\right)}{n} = \frac{\sum_{j=1}^{J}\left(\sum_{i=1}^{n_j}\beta_{ij}\right)}{n} = \frac{\sum_{i,j}\beta_{ij}}{n}$$

```
. mean beta

Mean estimation                         Number of obs    =     100

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        beta |   .0000533   .0000935     -.0001321    .0002388
--------------------------------------------------------------
```

Which Approach should we use…A or B?
Try an analysis of variance:

```
. oneway beta region
```

```
                    Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      2.2192e-06      6   3.6987e-07      0.41     0.8719
 Within groups      .000084249     93   9.0591e-07
------------------------------------------------------------------------
    Total           .000086469     99   8.7342e-07

Bartlett's test for equal variances:  chi2(6) =  30.5327  Prob>chi2 = 0.000
```

No evidence for difference in means of the regions since the F-statistic comparing the ratio of the $MS_{between}$ to the MS_{within} has a p-value of 0.87. According to the ANOVA we should use approach B. But…the ANOVA requires the assumption that the variance within region is the same for each region. We have shown this to be false with Bartlett's test for equal variances. Hence, we will try approach C a compromise that uses a weighted combination of approaches A and B.

Approach C: Weighted combination of A and B

Taking a short cut in stata, we specify a model with a random intercept for region, then obtain the empirical Bayes estimates for each region

There are three ways we can specify a random intercept for our continuous outcome

xtreg – doesn't work for our data since likelihood too difficult to maximize

```
. xtreg beta, re i(region) mle
```

xtmixed – equiv to xtreg and doesn't work for same reason

```
. xtmixed beta || region:, mle
```

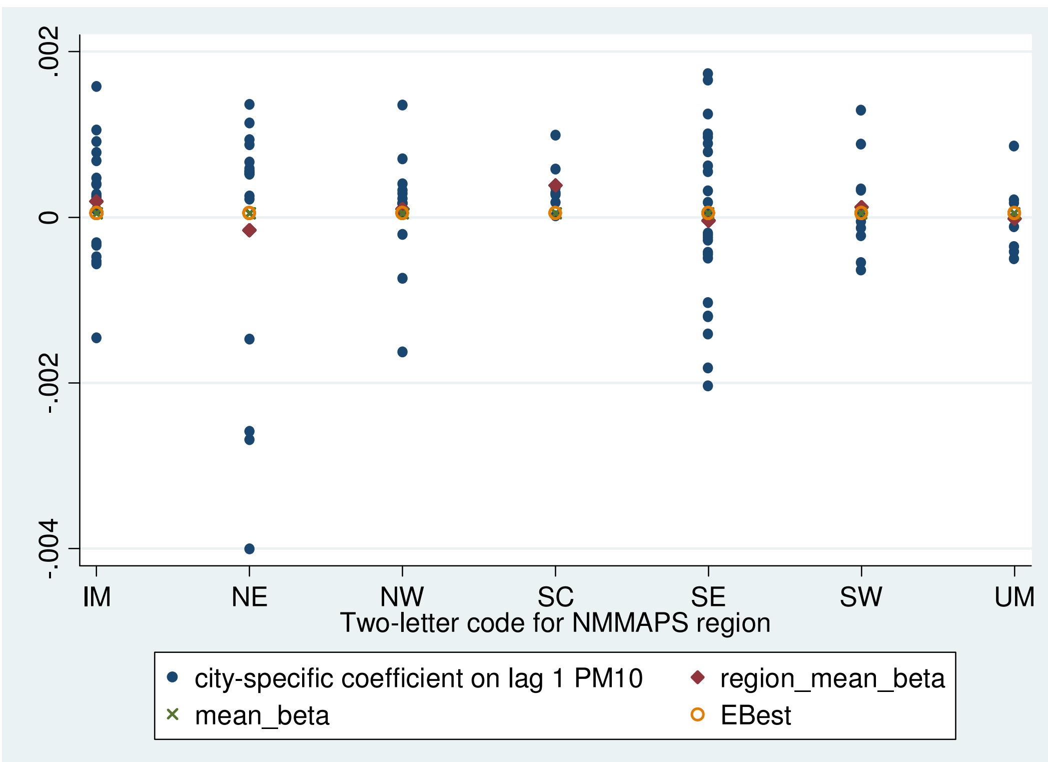
gllamm – works!

```
. gllamm beta, i(region) adapt nip(15)

Running adaptive quadrature
Iteration 0:    log likelihood =  504.47871
Iteration 1:    log likelihood =  507.63647
Iteration 2:    log likelihood =  516.66851
Iteration 3:    log likelihood =  526.03579
Iteration 4:    log likelihood =  535.76318
Iteration 5:    log likelihood =  545.78396
Iteration 6:    log likelihood =  554.51778
Iteration 7:    log likelihood =  554.60229
Iteration 8:    log likelihood =  554.78871
Iteration 9:    log likelihood =   555.8968
Iteration 10:    log likelihood =  556.14884
Iteration 11:    log likelihood =  556.15105
Iteration 12:    log likelihood =  556.15105


Adaptive quadrature has converged, running Newton-Raphson
Iteration 0:   log likelihood =  556.15105
Iteration 1:   log likelihood =  556.15105

number of level 1 units = 100
number of level 2 units = 7

Condition Number = 760.4234
```

```
gllamm model

log likelihood = 556.15105

------------------------------------------------------------------------------
        beta |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .0000532    .000093     0.57   0.567    -.0001291    .0002355
------------------------------------------------------------------------------

Variance at level 1
------------------------------------------------------------------------------
  8.650e-07 (1.224e-07)

Variances and covariances of random effects
------------------------------------------------------------------------------
***level 2 (region)

    var(1): 3.210e-16 (4.177e-12)
------------------------------------------------------------------------------
```

We interpret 8.647e-07 to be the estimated variance of the residuals and 3.210e-16 to be the estimated variance of the region-specific intercepts.

```
scatter beta region_mean_beta mean_beta EBest region, xlabel(1 2 3 4 5 6 7,
valuelabel) msymbol(o d X Oh)
```



Note that the EB estimates are weighted combinations of the region_mean_beta (Approach A) and the mean_beta (Approach B). Most of the EB estimates fall closer to the region_mean_beta except for the NE region, which falls closer to the mean_beta estimate.