Homework 09 Solutions

12/01/2019

2

We need to assume throughout this problem that the errors ϵ_i are independent of the quantities η_j. (Remember, ϵ_i represents the "unexplainable variation"; see slide 14 of lecture 23.)

(a)

Since, for two random variables A, B, we have that $\text{Corr}(A, B) = 0 \iff \text{Cov}(A, B) = 0$, let's just focus on the covariance. We have:

$$\begin{aligned} \text{Cov}(y_i, y_{i'}) &= \text{Cov}(\alpha_{j[i]} + \epsilon_i, \alpha_{j[i']} + \epsilon_{i'}) && (1) \\ &= \text{Cov}(\alpha_{j[i]}, \alpha_{j[i']}) + \text{Cov}(\alpha_{j[i]}, \epsilon_{i'}) + \text{Cov}(\epsilon_i, \alpha_{j[i']}) + \text{Cov}(\epsilon_i, \epsilon_{i'}) && (2) \end{aligned}$$

Recall that if two random variables A and B are independent, then $f(A)$ and $g(B)$ are independent for any functions f and g[1]. Since $\eta_{j[i]} \perp\!\!\!\perp \eta_{j[i']}$, and β_0 is just a constant, it follows that $\alpha_{j[i]} \perp\!\!\!\perp \alpha_{j[i']}$

Furthermore, we have the errors ϵ_i are independent of everything, by assumption.

Hence each pair of variables in (2) is independent, so the covariances are all 0.

(b)

Substituting $j[i]$ for $j[i']$ in (2), we have

$$\begin{aligned} \text{Cov}(y_i, y_{i'}) &= \text{Cov}(\alpha_{j[i]}, \alpha_{j[i]}) + \text{Cov}(\alpha_{j[i]}, \epsilon_{i'}) + \text{Cov}(\epsilon_i, \alpha_{j[i]}) + \text{Cov}(\epsilon_i, \epsilon_{i'}) && (3) \\ &= \text{Var}(\alpha_{j[i]}) + 0 + 0 + 0 && (4) \\ &= \tau^2 && (5) \end{aligned}$$

Now,

$$\text{Corr}(y_i, y_{i'}) = \frac{\text{Cov}(y_i, y_{i'})}{\sqrt{\sigma^2_{y_i} \sigma^2_{y_{i'}}}} \tag{6}$$

Writing the model in variance components form, we have $y_i = \beta_0 + \eta_{j[i]} + \epsilon_i$, which means $y_i \sim N(\beta_0, \sigma^2 + \tau^2)$ (since the sum of two independent normally distributed random variables has variance equal to the sum of the variances). Substituting this into (6) yields

$$\begin{aligned} \text{Corr}(y_i, y_{i'}) &= \frac{\text{Cov}(y_i, y_{i'})}{\sqrt{(\sigma^2 + \tau^2)(\sigma^2 + \tau^2)}} && (7) \\ &= \frac{\tau^2}{\sigma^2 + \tau^2} && (8) \end{aligned}$$

[1] Any measurable functions f and g, that is, but you don't really need to worry about this technicality.

(c)

$$\frac{1}{n_j}\sum Y_i = \frac{1}{n_j}[\sum \alpha_j + \sum \epsilon_i] \tag{9}$$

$$= \alpha_j + \frac{1}{n_j}\sum \epsilon_i \tag{10}$$

$$\implies \text{Var}(\bar{y}_j) = \text{Var}(\alpha_j) + \frac{1}{n_j^2}\sum \text{Var}(\epsilon_i) \tag{11}$$

$$= \tau^2 + \frac{\sigma^2}{n_j} \tag{12}$$

(d)

Notice that this is equivalent to sampling a group, sampling a bunch of data points from that group, splitting those data points into two sub-groups, and taking the means of each subgroup. We have:

$$\text{Cov}(\bar{y}_j, \bar{y}_j^*) = \text{Cov}\left(\frac{1}{n_j}\sum_i y_i, \frac{1}{n_j^*}\sum_k y_k^*\right) \tag{13}$$

$$= \frac{1}{n_j^2}\sum_i\sum_k \text{Cov}(y_i, y_k^*) \tag{14}$$

$$= \frac{1}{n_j^2} n_j^2 \tau^2 \tag{15}$$

$$= \tau^2 \tag{16}$$

where in line 2 we assumed that $n_j^* = n_j$, and in line 3 we used that $\text{Cov}(y_i, y_k^*)$ is equivalent to $\text{Cov}(y_i, y_{i'})$ from part (b) above. Now, using the result from part (c), we have

$$\text{Cov}(\bar{y}_j, \bar{y}_j^*) = \frac{\tau^2}{\sqrt{\sigma^2_{\bar{y}}\sigma^2_{\bar{y}^*}}} \tag{17}$$

$$= \frac{\tau^2}{\tau^2 + \sigma^2/n_j} \tag{18}$$

3

(a)

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: per.cap.income ~ pct.hs.grad + (pct.hs.grad | state)
##    Data: dat
##
## REML criterion at convergence: 8293
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -3.146 -0.583 -0.085  0.454  4.321
```

```
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  state    (Intercept) 7949288  2819.4
##           pct.hs.grad    3676    60.6   -1.00
##  Residual             8115992  2848.9
## Number of obs: 440, groups:  state, 48
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  -4920.7     1699.0    -2.9
## pct.hs.grad    297.2       23.1    12.9
##
## Correlation of Fixed Effects:
##             (Intr)
## pct.hs.grad -0.984
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

```
##        grp        var1        var2    vcov sdcor
## 1    state (Intercept)        <NA> 7949288  2819
## 2    state pct.hs.grad        <NA>    3676    61
## 3    state (Intercept) pct.hs.grad -170937    -1
## 4 Residual        <NA>        <NA> 8115992  2849
```

The parameter estimates are as follows.

Random effects estimates

$$\begin{aligned}
\widehat{\sigma^2} &= 8,115,992 \\
\widehat{\tau_0^2} &= 7,949,288 \\
\widehat{\tau_1^2} &= 3,676 \\
\widehat{\text{Corr}}(\eta_0, \eta_1) &= -0.99999
\end{aligned}$$

Fixed effects estimates

$$\begin{aligned}
\widehat{\beta_0} &= -4,921 \\
\text{SE}\left(\widehat{\beta_0}\right) &= 1,699 \\
\widehat{\beta_1} &= 297.23 \\
\text{SE}\left(\widehat{\beta_1}\right) &= 23.05
\end{aligned}$$

If we fit the model calling `lmer` and setting the option within `control = lmerControl(optimizer = 'bobyqa')` we get a similar but slightly different model.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: per.cap.income ~ pct.hs.grad + (pct.hs.grad | state)
##    Data: dat
## Control: lmerControl(optimizer = "bobyqa")
##
```

```
## REML criterion at convergence: 8292
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -3.053 -0.575 -0.096  0.453  4.244
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  state    (Intercept) 23008246 4797
##           pct.hs.grad     7397   86     -1.00
##  Residual              8064024 2840
## Number of obs: 440, groups:  state, 48
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  -4455.3     1791.2   -2.49
## pct.hs.grad    291.3       24.8   11.73
##
## Correlation of Fixed Effects:
##             (Intr)
## pct.hs.grad -0.987
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

(b)

See Figure 1.

(c)

Examining Figure 1 we see that the pooled regression and mlm β's agree almost exactly, which makes sense given that they are both models considering all of the data. There is a lot more variance in the unpooled regression lines since they are estimated individually for each state. We also note that for states with only one observation we are unable to estimate the state level regression. The mlm α's fall somewhere between the mlm β's and the unpooled regression lines. In states with very little data they agree with the mlm β's (e.g. HI, DE) while in states with a lot of data (e.g. CT, NJ) they agree more closely with the unpooled regression line.

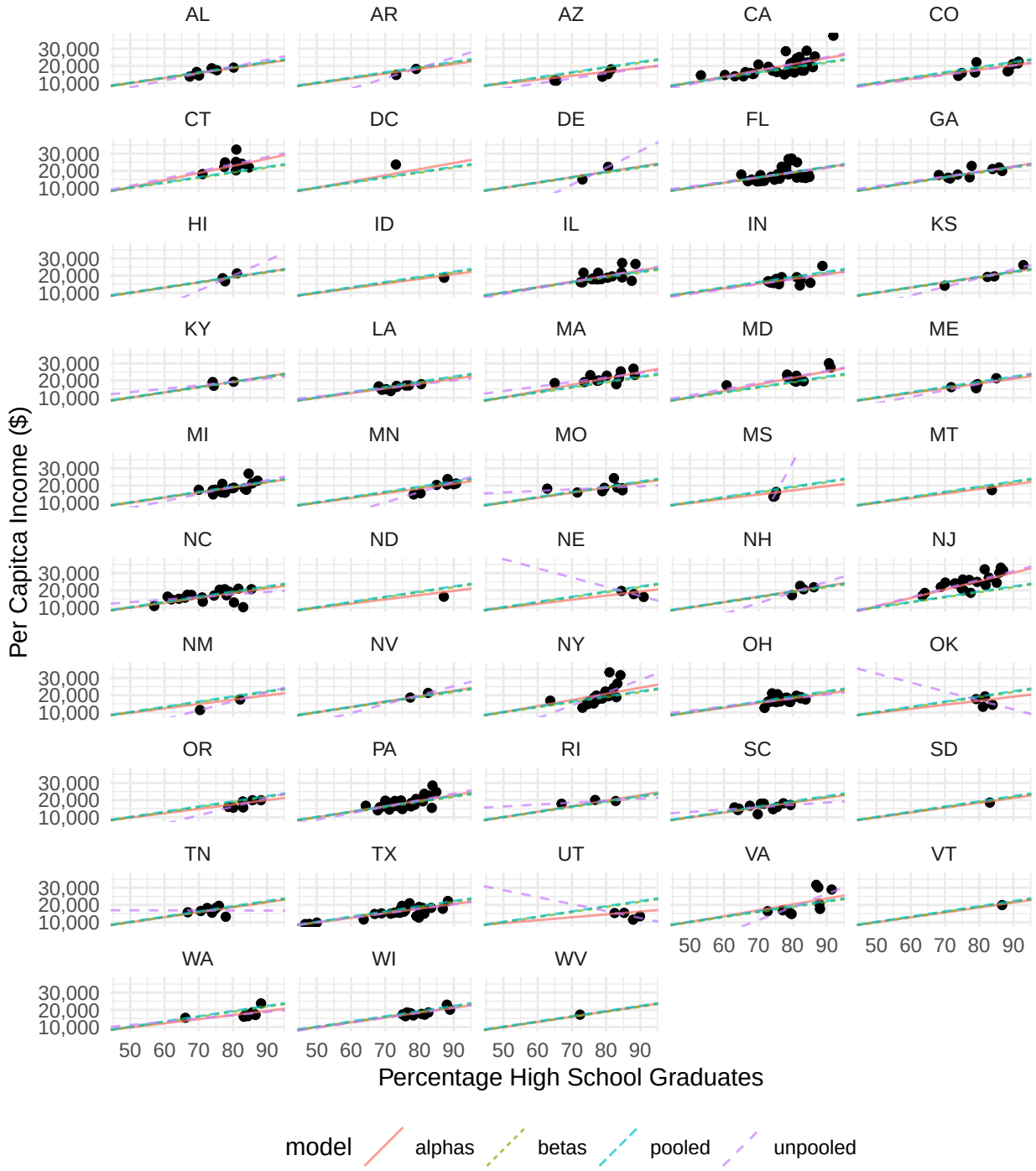
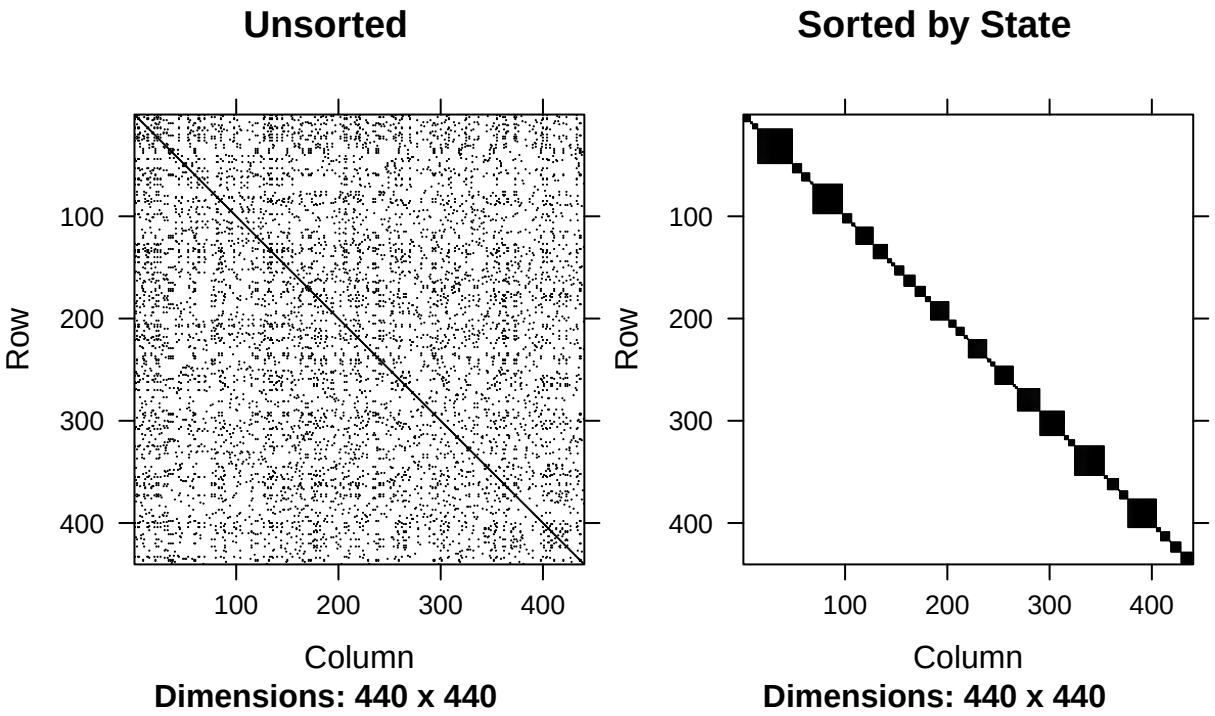


Figure 1: Problem 3b: HS graduation rate vs. county level per capita income with regression lines.

(d)



We plot the unsorted variance-covariance matrix in the left panel of Figure **??** and the same matrix with rows and columns sorted by state in the right panel. The right panel of Figure **??** makes is clear that the estimated covariance between counties in the same state is high and zero (by assumption) between counties in different states. Larger squares correspond to states with more observed counties.

(e)

QQ plots are shown in Figure 2 while the marginal residuals, conditional residuals, and random effects are shown in Figures 3, 4, and 5 respectively. The QQ plots all show a heavy right tail as well as a fewer large (in absolute value) residuals in the left tail. The marginal residuals generally appear mean zero but suggest that the deviations from normality may be coming primarily from CA and NJ which do not appear mean zero when considered individually. It is worth noting that these states both contain a lot of observations and have relatively large random effects which is what makes them so noticeable. The conditional residuals appear good for the most part but there may be some trend in PA although it is hard to know how to interpret this since they are plotted against an index and not a covariate. The random effects appear fine with the residuals clustered around the state mean in all cases.

Given the deviation from normality observed in the qq plots we should consider a transformation of `per.cap.income`, perhaps a log transform, to try to address the heavy tails. The conditional residuals suggest that we might be able to make some improvements in how the random effects are modeled but that these are likely to be marginal improvements affecting only a few states.

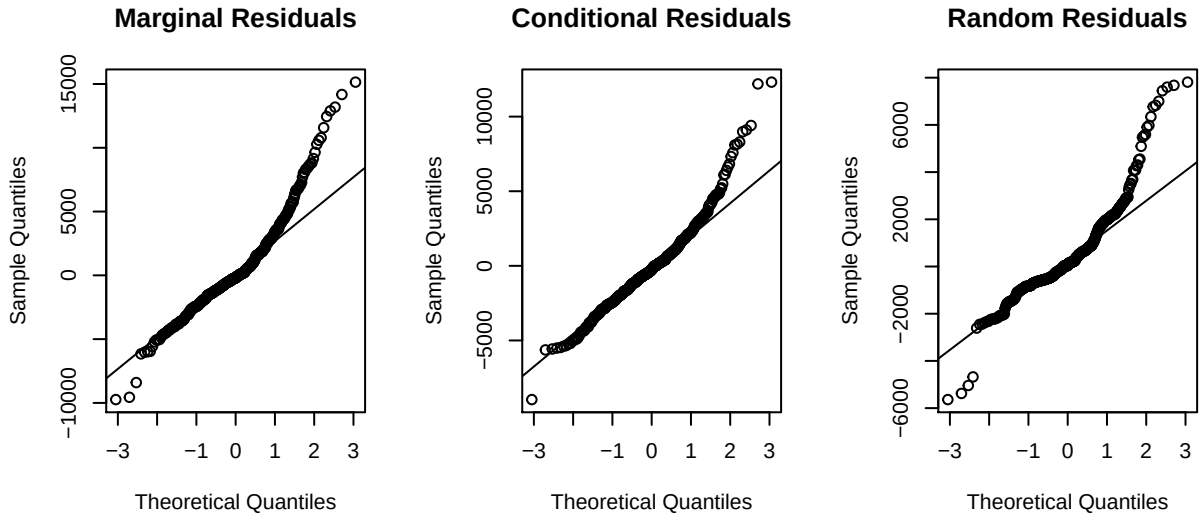


Figure 2: Problem 3e: QQ plot of multilevel model residuals.

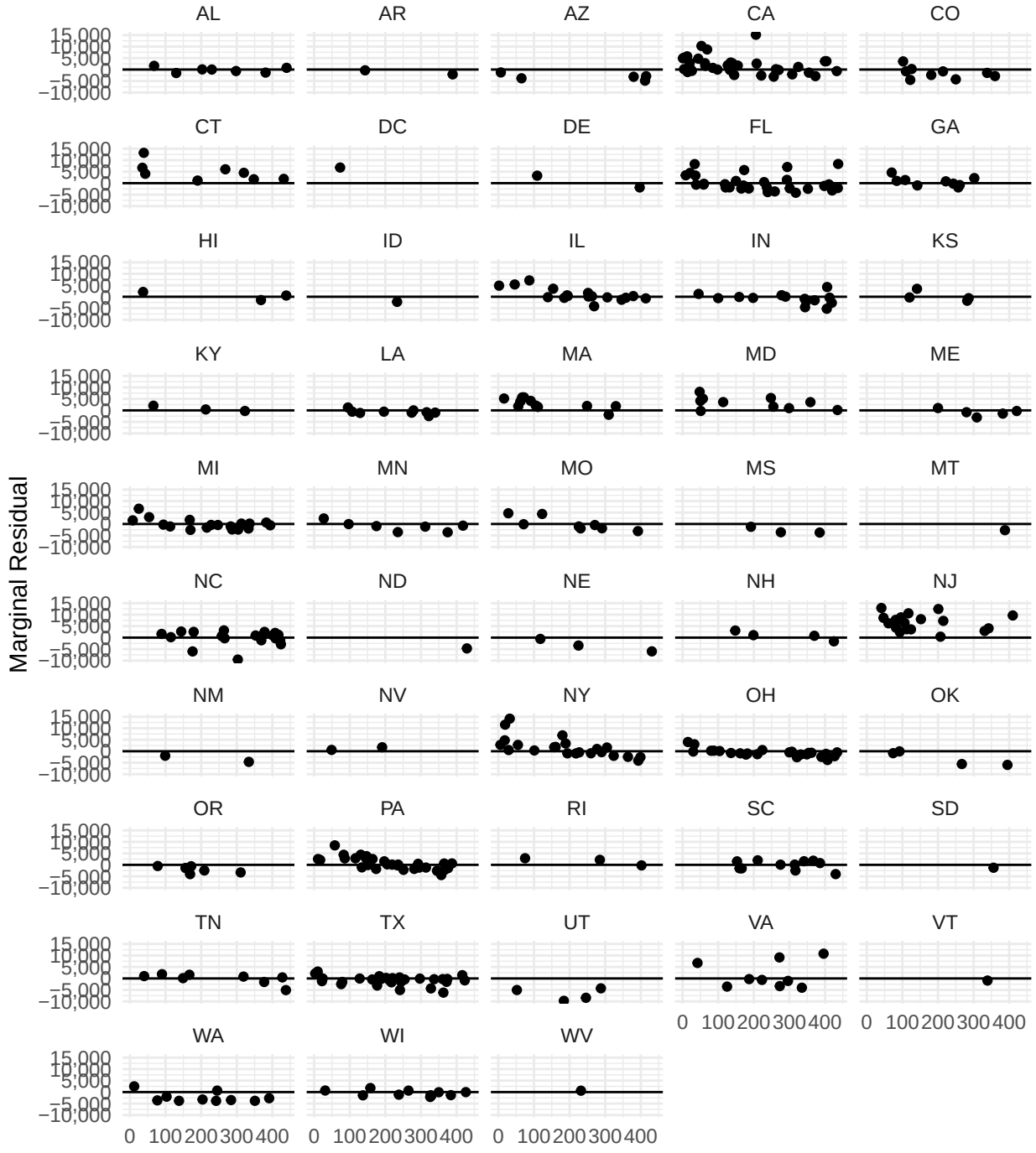


Figure 3: Problem 3e: Facet plot of marginal residuals.

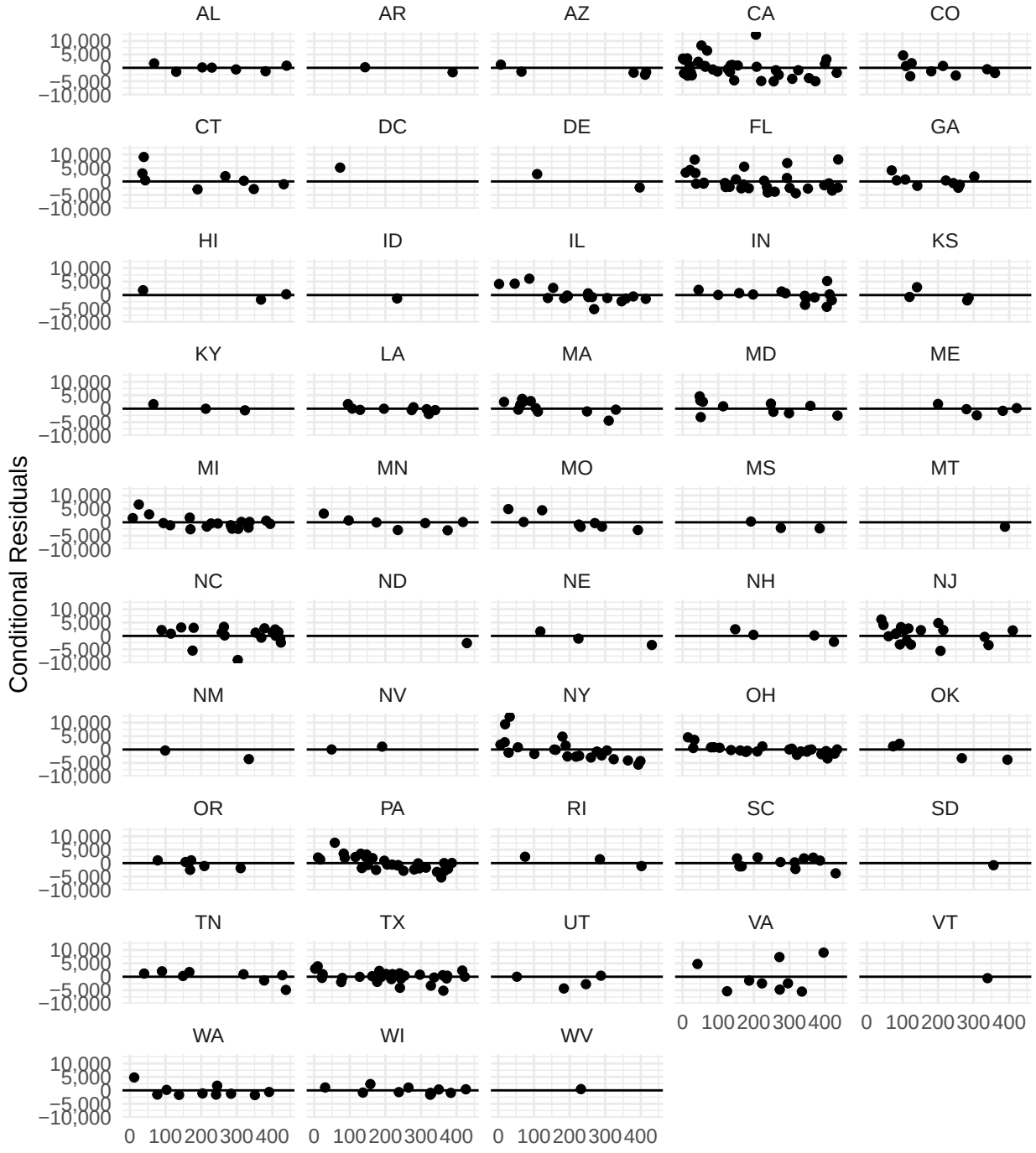


Figure 4: Problem 3e: Facet plot of conditional residuals.

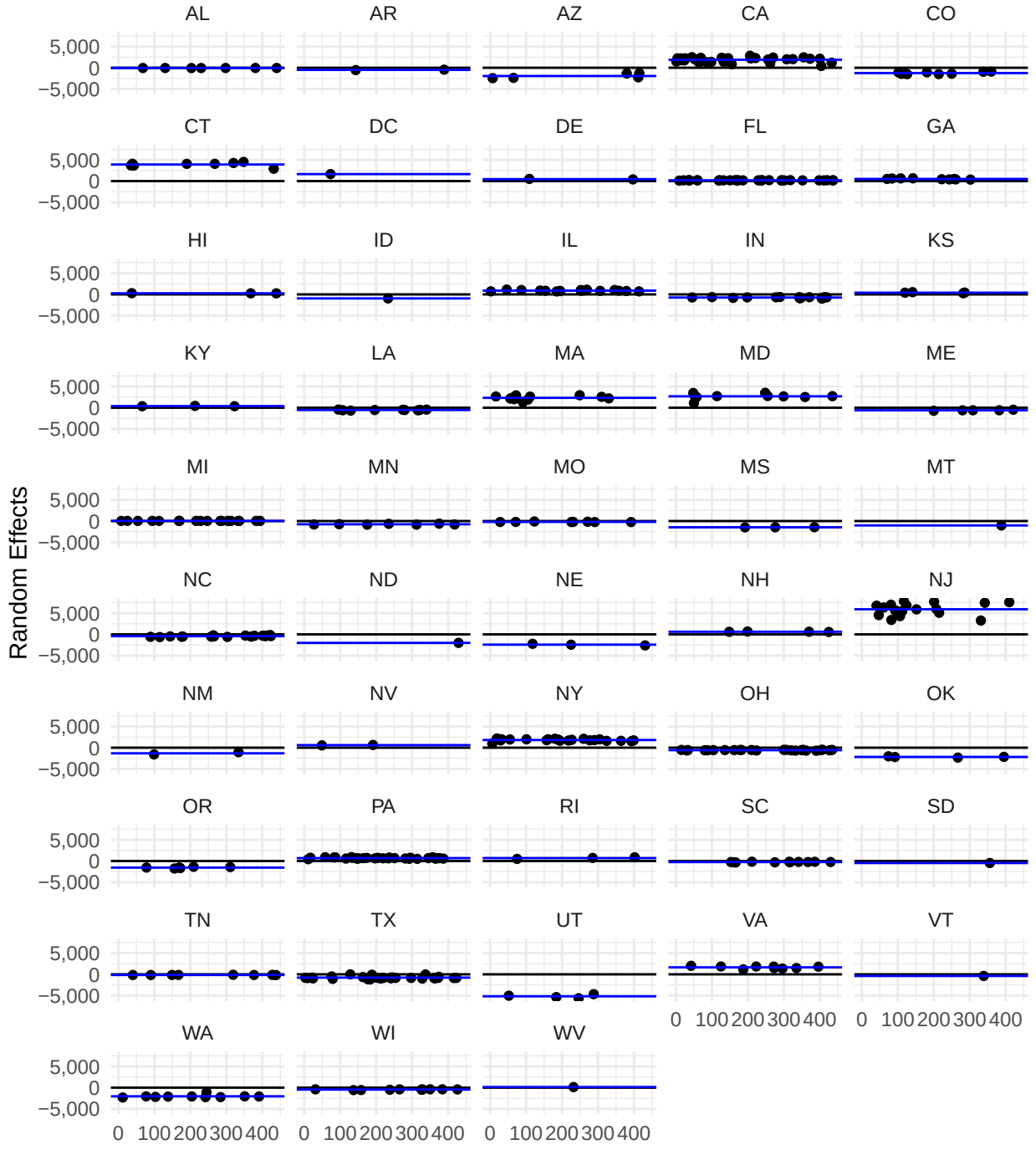


Figure 5: Problem 3e: Facet plot of random effects. Blue line shows mean state random effect.