MATH TEST DATASET

dataset summary:

- $p = 100$ highschools
- $n_i = \#$ student for $i = 1 \dots p$
  ↳ $n_i \geq 400 \quad \forall i$
- y_{ij} = score for jth student in school i

$$y_{ij} \sim N(\theta_i, \sigma^2) \quad \forall i = 1 \dots p, \ \forall j = 1 \dots n_j$$

model:

$$\theta_i \sim N(\mu, \tau^2 \sigma^2) \quad \forall i = 1 \dots p$$

A) Find MLE for θ_i, $\forall i = 1 \dots p$

$$\mathcal{L}(\theta, y) = \prod_{i=1}^{p} \prod_{j=1}^{n_i} p(y_{ij} | \theta_i)$$

$$\propto \prod_{i=1}^{p} \prod_{j=1}^{n_i} \exp\left\{ -\frac{1}{2\sigma^2} (y_{ij} - \theta_i)^2 \right\}$$

$$= \prod_{i=1}^{p} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n_i} (y_{ij} - \theta_i)^2 \right\}$$

$$= \prod_{i=1}^{p} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n_i} (y_{ij}^2 - 2y_{ij}\theta_i + \theta_i^2) \right\}$$

$$\propto \prod_{i=1}^{p} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n_i} (\theta_i^2 - 2y_{ij}\theta_i) \right\}$$

so that

$$\log \mathcal{L}(\theta, y) \propto \sum_{i=1}^{p} \frac{1}{2\sigma^2} \sum_{j=1}^{n_i} (\theta_i^2 - 2y_{ij}\theta_i)$$

so that

$$\frac{\partial \mathcal{L}(\theta, y)}{\partial \theta_i} \propto \sum_{j=1}^{n_i} (\theta_i - y_{ij}) = n_i\theta_i - \sum_{j=1}^{n_i} y_{ij}$$

so that

$$\frac{\partial \mathcal{L}(\theta, y)}{\partial \theta_i} = 0 \longrightarrow \hat{\theta}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \overline{y_i} \quad , \forall i \in \{1 \dots p\}$$

B) As suggested, we can notice that we observe extreme score (average) in schools in which n_i is higher. This is because smaller datasets will be more affected by outliers in their estimate.

(see notebook)

c) Consider the following two level hierarchical model:

- $y_{ij} \mid \theta_i, \sigma^2 \sim N(\theta_i, \sigma^2)$
- $\theta_i \mid \tau^2, \sigma^2 \sim N(\mu, \tau^2\sigma^2)$

<u>with priors:</u>

- $p(\mu) \propto c$
- $p(\sigma^2) \sim$ jeffrey's $\longrightarrow p(\sigma^2) = \dfrac{1}{\sigma^2}$
- $\tau^2 \sim \text{In-Ga}\left(\frac{1}{2}, \frac{1}{2}\right)$

In order to fit a Gibbs sampler for the parameters, we first need to derive the full conditionals:

1) $p(\theta_i \mid y, \sigma^2, \mu, \tau^2)$
2) $p(\sigma^2 \mid \theta, y, \mu, \tau^2)$
3) $p(\mu \mid \theta, \sigma^2, y, \tau^2)$
4) $p(\tau^2 \mid \theta, \sigma^2, y, \mu)$

1) $p(\theta_i|-) \propto \prod_{j=1}^{n_i} p(y_{ij}|\theta_i,\sigma^2)\cdot p(\theta_i)$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\sum^{n_i}(y_{ij}-\theta_i)^2\right\}\cdot\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i-\mu)^2\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\sum^{n_i}(\theta_i^2-2y_{ij}\theta_i)\right\}\cdot\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i^2-2\theta_i\mu)\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}(n_i\theta_i^2-2\bar{y}_i\cdot n_i\theta_i)\right\}\cdot\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i^2-2\theta_i\mu)\right\}$$

$$= \exp\left\{-\frac{n_i}{2\sigma^2\tau^2}\left(\theta_i^2\tau^2-2\bar{y}_i\tau^2\cdot\theta_i + \frac{\theta_i^2}{n_i}-\frac{2\theta_i\mu}{n_i}\right)\right\}$$

$$= \exp\left\{-\frac{n_i}{2\sigma^2\tau^2}\left(\theta_i^2(\tau^2+n_i^{-1})-2\theta_i\left(\bar{y}_i\cdot\tau^2+\mu\cdot n_i^{-1}\right)\right)\right\}$$

$$= \exp\left\{-\frac{n_i(\tau^2+n_i^{-1})}{2\sigma^2\tau^2}\left(\theta_i^2-2\theta_i\left(\frac{\bar{y}_i\cdot\tau^2+\mu\cdot n_i^{-1}}{\tau^2+n_i^{-1}}\right)\right)\right\}$$

completing the square:

$$\theta_i|- \sim N\left(\frac{\bar{y}_i\cdot\tau^2+\mu\cdot n_i^{-1}}{\tau^2+n_i^{-1}}\ ;\ \frac{\sigma^2\tau^2}{(n_i\tau^2+1)}\right)$$

⇒ Note: we can update all θ_i together using

$\theta|- \sim MVN(m, V)$ (1)

with:
- m = vectors of means as in
- V = diagonal matrix with elements as in

2) $p(\sigma^2|-) \propto \left(\prod_{i=1}^{p}\left(\prod_{j=1}^{n_i} p(y_{ij}|\theta_i,\sigma^2)\right) p(\theta_i|\sigma^2,\tau^2,\mu)\right) p(\sigma^2)$

$$\propto \left(\prod_{i=1}^{p}\left(\prod_{j=1}^{n_i}\frac{1}{\sigma}\cdot\exp\left\{-\frac{1}{2\sigma^2}(y_{ij}-\theta_i)^2\right\}\right)\cdot\frac{1}{\sigma}\cdot\exp\left\{-\frac{1}{2\sigma^2\tau^2}(\theta_i-\mu)^2\right\}\right)\frac{1}{\sigma^2}$$

matrix form for θ

$$= (\sigma^2)^{-\frac{N+p}{2}-1}\cdot\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2+(\theta-\mu)'(\tau^2 I)^{-1}(\theta-\mu)\right)\right\}$$

so that:

$$\sigma^2|- \sim IG\left(\frac{\overbrace{\sum^{p} n_i}^{N}+p}{2};\ \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2+(\theta-\mu)'(\tau^2 I)^{-1}(\theta-\mu)}{2}\right) \quad (2)$$

3) $p(\mu|-) \propto \left(\prod_{i=1}^{p} p(\theta_i|\mu,\sigma^2,\tau^2)\right)\cdot p(\mu)$

$$\propto \exp\left\{-\frac{1}{2\sigma^2\tau^2}\sum_{i=1}^{p}(\theta_i-\mu)^2\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2\tau^2}\left(p\mu^2-2p\bar{\theta}\cdot\mu\right)\right\}$$

$$= \exp\left\{-\frac{p}{2\sigma^2\tau^2}\left(\mu^2-2\bar{\theta}\mu\right)\right\}$$

completing the square:

$$\mu|- \sim N\left(\bar{\theta},\ \frac{\sigma^2\tau^2}{p}\right) \quad (3)$$

4) $p(\tau^2 \mid -) \propto \left(\prod_{i=1}^{p} p(\theta_i \mid \mu, \delta^2, \tau^2)\right) p(\tau^2)$

$$\propto \left(\frac{1}{\tau}\right)^{p} \cdot \exp\left\{-\frac{1}{2\delta^2\tau^2}\sum_{i=1}^{p}(\theta_i-\mu)^2\right\} \cdot (\tau^2)^{-\frac{1}{2}-1} \cdot \exp\left\{-\frac{1}{2}\tau^2\right\}$$

$$= (\tau^2)^{-\frac{p}{2}-\frac{1}{2}-1} \cdot \exp\left\{-\frac{1}{\tau^2}\left(\frac{(\theta-\mu)'(\delta^2 I)^{-1}(\theta-\mu)+1}{2}\right)\right\}$$

using matrix notation for θ

so that

$$\tau^2 \mid - \sim \text{In-Ga}\left(\frac{p+1}{2}, \frac{(\theta-\mu)'(\delta^2 I)^{-1}(\theta-\mu)+1}{2}\right) \qquad (4)$$

Hence, a Gibbs-sampler will iteratively update the parameters using

(1), (2), (3), (4)

D) We have, as derived in c):

$$\theta_i \mid - \sim N\left(\frac{\bar{y}_i \cdot \tau^2 + \mu \cdot n_i^{-1}}{\tau^2 + n_i^{-1}} ; \frac{\sigma^2 \tau^2}{(n_i \tau^2 + 1)} \right)$$

so that

$$E(\theta_i \mid -) = \frac{\bar{y}_i \cdot \tau^2 + \mu \cdot n_i^{-}}{\tau^2 + n_i^{-}}$$

which can be expressed as a convex combination of prior mean and data mean. Indeed:

$$= \bar{y}_i \underbrace{\frac{\tau^2}{\tau^2 + n_i^{-1}}}_{1 - k_i} + \mu \underbrace{\frac{n_i^{-1}}{\tau^2 + n_i^{-1}}}_{k_i} \rightarrow \text{SHRINKAGE COEFFICIENT}$$

(see notebook for the plot)

One can notice how, when the group size (sample size) increases, the shrinkage effect is more evident.

⇒ Note that you can also write the model as:

$$y_{ij} = \mu + \delta_i + e_{ij}$$

$\delta_i \to N(0, \tau^2\sigma^2)$ $\qquad$ $e_{ij} \to N(0, \sigma^2)$

so that: $\qquad$ CONDITIONALLY ON μ:

- $\mathrm{cov}(y_{ij}; y_{ik})$ for $j \neq k$ (same school but different individual)

$$= E\left[(\mu + \delta_i + e_{ij} - E(y_{ij}))(\mu + \delta_i + e_{ij} - E(y_{ij}))\right]$$

$$= E\left[(\cancel{\mu} + \delta_i + e_{ij} - \cancel{\mu})(\cancel{\mu} + \delta_i + e_{ij} - \cancel{\mu})\right]$$

$$= E\left[(\delta_i^2 + e_{ij}\delta_i + e_{ik}\delta_i + e_{ij}\cdot e_{ik})\right]$$

$$= \underbrace{E(\delta_i^2)}_{\substack{\tau^2\sigma^2 \\ = VAR(\delta_i)}} + 2\underbrace{E(e_{ij}\delta_i)}_{\substack{= cov(e_{ij}, \delta_i) \\ = 0}} + \underbrace{E(e_{ij}\cdot e_{ik})}_{= 0 \text{ since independent}} = \boxed{\tau^2\sigma^2}$$

- $\mathrm{cov}(y_{ij}, y_{i'k})$ for $i \neq i'$

$$= E\left[(\delta_i + e_{ij})(\delta_{i'} + e_{i'k})\right]$$

$$= E\left[(\delta_i\cdot\delta_{i'} + e_{ij}\delta_{i'} + e_{i'k}\cdot\delta_i + e_{ij}\cdot e_{i'k})\right]$$

$$= \underbrace{E(\delta_i\cdot\delta_{i'})}_{\substack{= cov(\delta_i, \delta_{i'}) \\ = 0}} + \underbrace{E(e_{ij}\delta_{i'})}_{\substack{= 0 \\ (\text{indep.})}} + \underbrace{E(e_{i'k}\delta_i)}_{= 0} + \underbrace{E(e_{ij}\cdot e_{i'k})}_{\substack{\text{indep by hp.} \\ = 0}} = \boxed{\emptyset.}$$

If we consider assuming that observations coming from different schools were completely independent, I would say that I am not convinced by the assumption.
Indeed, think for example about the fact that all the observations come from the same country, the US, so some sort of correlation between different schools should be expected.

In practice, we are not making this assumption.
Indeed, we are assuming independence once conditioning on the ground mean μ, which could account for the common effect.

Hence, the assumption seems realistic.

f) We now focus on the assumption that σ^2 is the same for all schools.
This assumption doesn't allow for schools to differentiate in terms of variety of students' result, even when the number of students is considerably different.

We could for example have an "elite" school in which all students were really well prepared for the math exam and also have a school in which people in some classes focused more on math and scored higher while people in other classes focused more on classical studies and scored lower.

The model should be such that different school variances are allowed.

A more formal testing strategy for this hypothesis is presented in the code.

The strategy proposed in class was to perform a POSTERIOR PREDICTIVE CHECK.

In other words, we could simulate a dataset obtained using the posterior parameters and see if such a dataset "resembles" the original one.

For implementation, see mathscores notebook looking at the p-value obtained doing an F test for the standard deviations, the null hypothesis is not rejected and the assumption made by the model seems to be confirmed.

BLOOD PRESSURE DATASET

20 individuals → some are given a treatment and some are not.

↓

each with repeated observations about blood pressure

A) Want to check if the treatment is effective in reducing blood pressure.

⇒ Naive solution: perform a t-test on the difference of the means (taking all the observation from same patient as independent)

To perform this test is not a good strategy because it assumes that the sample are independent.
This is not the case: values are strongly correlated since they are longitudinal data coming from few individuals

(see notebook for the test)

B) The standard error of the test is considerably higher because now we are considering fewer data points (by collapsing data from the same individual to only be represented by the mean of that individual). (see notebook)

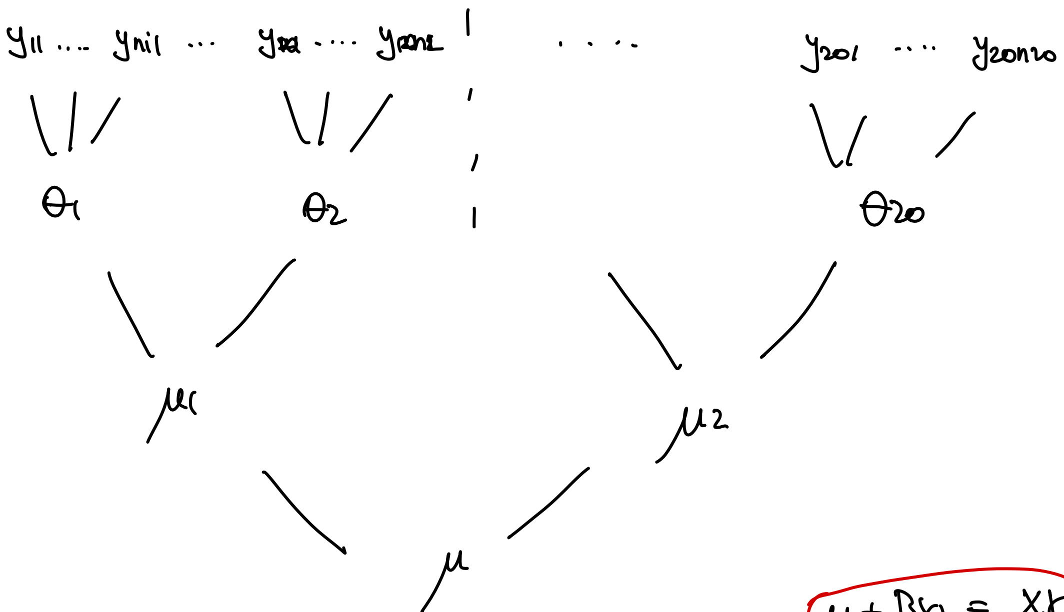
Still, this test is not ideal because of heteroskedasticity.

20 different subjects → groups : θ_i

γ

$y_{ij} \mid \theta_i, \delta^2 \sim N(\theta_i, \delta^2)$

$\theta_i \mid \tau^2, \delta^2, \beta \sim N(\mu + \beta x_1, \tau^2 \delta^2)$



$\mu + \beta x_1 = X\gamma$

__Priors:__

- $p(\mu, \beta) \propto c \rightarrow p(\gamma) \propto c$
- $p(\delta^2) \propto \frac{1}{\delta^2}$
- $p(\tau^2) \propto \text{Inv-Ga}\left(\frac{1}{2}, \frac{1}{2}\right)$

So that the derivations for the posterior will be similar:

1) $p(\theta_i|-) \propto \left(\prod^{n_i} p(y_{ij}|\theta_i,\delta^2)\right) p(\theta_i|\gamma,\delta^2\tau^2)$

$$\propto \left(\prod_{j=1}^{n_i} \exp\left\{-\frac{1}{2\delta^2}(y_{ij}-\theta_i)^2\right\}\right) \exp\left\{-\frac{1}{2\delta^2\tau^2}(\theta_i - x_i'\gamma)^2\right\}$$

will be the same of previous

$$= \exp\left\{-\frac{1}{2\delta^2}\left(\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2 + \tau^2(\theta_i - x_i'\gamma)^2\right)\right\}$$

and following the same steps followed in A) for the previous dataset, but with now $\mu = x_i'\mu$:

$$\Rightarrow \theta_i|- \sim N\left(\frac{\bar{y}_i \cdot \tau^2 + x_i'\gamma\, n_i^-}{\tau^2 + n_i^-}; \frac{\delta^2\tau^2}{(n_i\tau^2+1)}\right)$$

2) $p(\delta^2|-) \propto \prod_{i=1}^{p}\left(\prod_{j=1}^{n_i} p(y_{ij}|\theta_i,\delta^2)\right) \cdot p(\theta_i|\delta^2,\tau^2,\gamma) \cdot p(\delta^2)$

$$\left(\prod_{i=1}^{p}\left(\prod_{j=1}^{n_i}\frac{1}{\delta}\cdot\exp\left\{-\frac{1}{2\delta^2}(y_{ij}-\theta_i)^2\right\}\right)\cdot\frac{1}{\delta}\cdot\exp\left\{-\frac{1}{2\tau^2}(\theta_i - x_i'\gamma)^2\right\}\right)\frac{1}{\delta^2}$$

vector of 20

$$= (\delta^2)^{-\frac{P+N}{2}-1} \cdot \exp\left\{-\frac{1}{2\delta^2}\left(\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2 + (\theta - X\gamma)'\tau^2 I(\theta - X\gamma)\right\}\right.$$

$$\Rightarrow \delta^2|- \sim \text{In-Ga}\left(\frac{N+P}{2}; \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2 + (\theta - X\gamma)'\tau^2 I(\theta - X\gamma)}{2}\right)$$

3) γ → now bidimensional. Still we are giving flat prior

$$p(\gamma|-) \propto \prod_{i=1}^{p} p(\theta_i|\gamma, \sigma^2, \tau^2) \cdot p(\gamma)$$

θ_i conditional indep

$$\propto \exp\left\{-\frac{1}{2}(\theta - X\gamma)'(\tau^2\sigma^2 I)^{-1}(\theta - X\gamma)\right\}$$

⇒ Now looks more like the posterior you would have under a linear model

$y - X\beta$

$$\propto \exp\left\{-\frac{1}{2}(\theta - X\gamma)'\Lambda(\theta - X\gamma)\right\}$$

$$= \exp\left\{-\frac{1}{2}\left[\theta'\Lambda\theta - \gamma'X'\Lambda\theta + \gamma'X'\Lambda X\beta - \theta'\Lambda X\beta\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left[-2\gamma'X'\Lambda\theta + \gamma'X'\Lambda X\gamma\right]\right\}$$

so that completing the square:

$\gamma|- \sim N(m, V)$

- $V = (X'\Lambda X)^{-1} = (X'(\sigma^2\tau^2 I)^{-1}X)^{-1} = \sigma^2\tau^2(X'X)^{-1}$
- $m = \cancel{\sigma^2\tau^2}(X'X)^{-1} \cdot X' \frac{1}{\cancel{\sigma^2\tau^2}} I\theta = (X'X)^{-1}X'\theta$

4) $$p(\tau^2|-) \propto \left(\prod_{i=1}^{p} p(\theta_i|\gamma, \sigma^2, \tau^2)\right) p(\tau^2)$$

$$\propto \left(\frac{1}{\tau^2}\right)^{p} \cdot \exp\left\{-\frac{1}{2\tau^2}(\theta - X\gamma)'(\sigma^2 I)^{-1}(\theta - X\gamma)\right\} \cdot (\tau^2)^{-\frac{1}{2}-1} \cdot \exp\left\{-\frac{1}{2}\tau^2\right\}$$

$$= (\tau^2)^{-p-\frac{1}{2}-1} \cdot \exp\left\{-\frac{1}{\tau^2}\left(\frac{(\theta - X\gamma)'(\sigma^2 I)^{-1}(\theta - X\gamma) + 1}{2}\right)\right\}$$

$$\Rightarrow \tau^2|- \propto \text{In-Ga}\left(p + \frac{1}{2}, \frac{(\theta - X\gamma)'(\sigma^2 I)^{-1}(\theta - X\gamma) + 1}{2}\right\}$$

D) (see notebook)

looking at autocorrelation plot, the hypotesis of independence (once conditioned on θ_i), seems to be realistic