A. Show (somewhat trivially) that the maximum likelihood estimate for θ is just the vector of sample means $\hat{\theta}_{MLE} = (\bar{y}_1, ..., \bar{y}_P)$.

First, we get the likelihood and log-likelihood:

$$L(\theta) = \prod_{i=1}^{P} \prod_{j=1}^{N_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_{ij} - \theta_i)^2\right\}$$

$$\begin{aligned}\mathcal{L}(\theta) &= \log\left[\prod_{i=1}^{P} \prod_{j=1}^{N_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_{ij} - \theta_i)^2\right\}\right] \\ &= \sum_{i=1}^{P} \sum_{j=1}^{N_i} \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_{ij} - \theta_i)^2\right]\end{aligned}$$

From here, we can maximize the log-likelihood to obtain the MLE for $\theta = (\theta_1, ..., \theta_P)$:

$$\begin{aligned}\frac{\partial}{\partial\theta_i}\mathcal{L}(\theta_i) &= \sum_{j=1}^{N_i} \frac{1}{\sigma^2}(y_{ij} - \theta_i) \\ &\stackrel{set}{=} 0 \\ \hat{\theta}_{MLE} &= \frac{\sum_{j=1}^{N_i} y_{ij}}{N_i} = \bar{y}_i\end{aligned}$$

B. Make a plot that illustrates the following fact: extreme school-level averages $\bar{y}_i$ (both high and low) tend to be at schools where fewer students were sampled. Explain briefly why this would be.

In Figure 1, we see that at schools with a smaller student population (i.e., smaller sample size), the in-school test scores have a large variance; consequently, the school-level averages are more variable. However, as school population increases, we see a regression towards mediocrity in the sense that school-level averages display smaller variance.

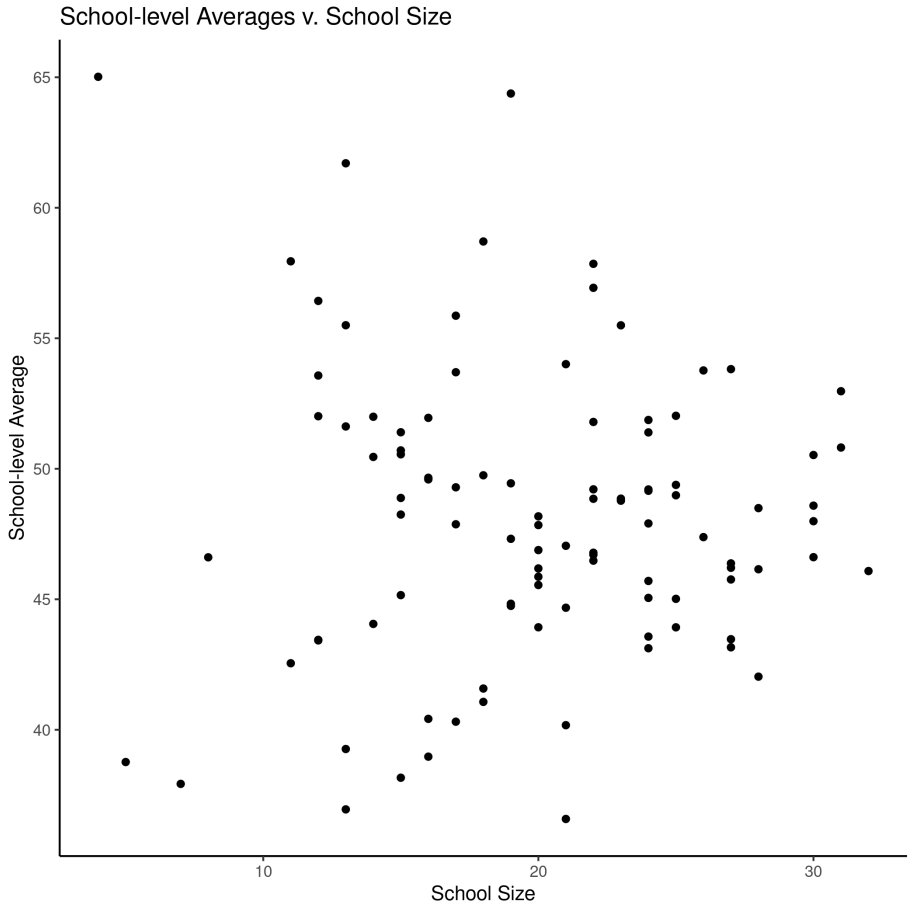

Figure 1: School-level Averages v. Sample Size

C. Fit the following two-level hierarchical model to these data via Gibbs sampling:

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

As a staring point, use a flat prior on μ, Jeffreys' prior on σ^2, and an IG$(1/2, 1/2)$ prior on τ^2.

Before we fit the model via Gibbs sampler, we need to obtain the full-conditionals for $\mu, \tau^2, \sigma^2, \theta$:

$$\begin{aligned}
p(\mu|\cdot) &\propto \prod_{i=1}^{P} p(\theta_i|\mu, \tau^2\sigma^2) \cdot p(\mu) \\
&\propto \prod_{i=1}^{P} \exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i - \mu)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\tau^2\sigma^2}\sum_{i=1}^{P}(\theta_i^2 - 2\mu\theta_i + \mu^2)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-2\left(\frac{\sum_{i=1}^{P}\theta_i}{\tau^2\sigma^2}\right)\mu + \mu^2\left(\frac{P}{\tau^2\sigma^2}\right)\right]\right\}, \\
p(\theta_i|\cdot) &\propto \prod_{j=1}^{N_i}\left[\exp\left\{-\frac{1}{2\sigma^2}(y_{ij} - \theta_i)^2\right\}\right]\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i - \mu)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{N_i}(-2y_{ij}\theta_i + \theta_i^2)\right\}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i^2 - 2\mu\theta_i)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-2\left(\frac{\sum_{j=1}^{N_i}y_{ij}}{\sigma^2} + \frac{\mu}{\tau^2\sigma^2}\right)\theta_i + \theta_i^2\left(\frac{N_i}{\sigma^2} + \frac{1}{\tau^2\sigma^2}\right)\right]\right\}
\end{aligned}$$

so, we see that

$$\begin{aligned}
p(\mu|\cdot) &\equiv N\left(\frac{\sum_{i=1}^{P}\theta_i}{P}, \frac{\tau^2\sigma^2}{P}\right) \\
p(\theta_i|\cdot) &\equiv N\left(\left[\frac{N_i}{\sigma^2} + \frac{1}{\tau^2\sigma^2}\right]^{-1}\left(\frac{\sum_{j=1}^{N_i}y_{ij}}{\sigma^2} + \frac{\mu}{\tau^2\sigma^2}\right), \left[\frac{N_i}{\sigma^2} + \frac{1}{\tau^2\sigma^2}\right]^{-1}\right)
\end{aligned}$$

Next, let's get those full-conditionals for σ^2 and τ^2:

$$\begin{aligned}
p(\sigma^2|\cdot) &\propto \prod_{i=1}^{P}\left[\prod_{j=1}^{N_i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y_{ij}-\theta_i)^2\right\}\cdot\frac{1}{\sqrt{2\pi\tau^2\sigma^2}}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i-\mu)^2\right\}\right]\cdot\frac{1}{\sigma^2} \\
&\propto \sigma^{-2}\prod_{i=1}^{P}\left[\sigma^{-N_i}\exp\left\{-\frac{1}{\sigma^2}\sum_{j=1}^{N_i}\left(\frac{y_{ij}-\theta_i}{2}\right)^2\right\}\cdot\sigma^{-1}\exp\left\{-\frac{1}{\sigma^2}\left(\frac{(\theta_i-\mu)^2}{2\tau^2}\right)\right\}\right] \\
&\propto (\sigma^2)^{-\left(\frac{\sum_{i=1}^{P}\frac{N_i+P}{2}}{2}\right)-1}\exp\left\{-\frac{1}{\sigma^2}\left[\sum_{i=1}^{P}\sum_{j=1}^{N_i}\left(\frac{(y_{ij}-\theta_i)^2}{2}\right)+\sum_{i=1}^{P}\left(\frac{(\theta_i-\mu)^2}{2\tau^2}\right)\right]\right\}, \\
p(\tau^2|\cdot) &\propto \prod_{i=1}^{P}\left[\frac{1}{\sqrt{2\pi\tau^2\sigma^2}}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i-\mu)^2\right\}\right]\cdot(\tau^2)^{-\frac{1}{2}-1}\exp\left\{-\frac{1}{2\tau^2}\right\} \\
&\propto (\tau^2)^{-\left(\frac{P+1}{2}\right)-1}\exp\left\{-\frac{1}{\tau^2}\left[\frac{\sum_{i=1}^{P}(\theta_i-\mu)^2}{2\sigma^2}+\frac{1}{2}\right]\right\},
\end{aligned}$$

which means that we have the following full-conditionals for σ^2 and τ^2:

$$\begin{aligned}
p(\sigma^2|\cdot) &\equiv \text{IG}\left(\frac{\sum_{i=1}^{P}N_i+P}{2}, \sum_{i=1}^{P}\sum_{j=1}^{N_i}\left(\frac{(y_{ij}-\theta_i)^2}{2}\right)+\sum_{i=1}^{P}\left(\frac{(\theta_i-\mu)^2}{2\tau^2}\right)\right), \\
p(\tau^2|\cdot) &\equiv \text{IG}\left(\frac{P+1}{2}, \frac{\sum_{i=1}^{P}(\theta_i-\mu)^2}{2\sigma^2}+\frac{1}{2}\right)
\end{aligned}$$

Although this portion of the question does not directly ask for any graphics, I figured that I'd post my trace plots here, which can be seen below.

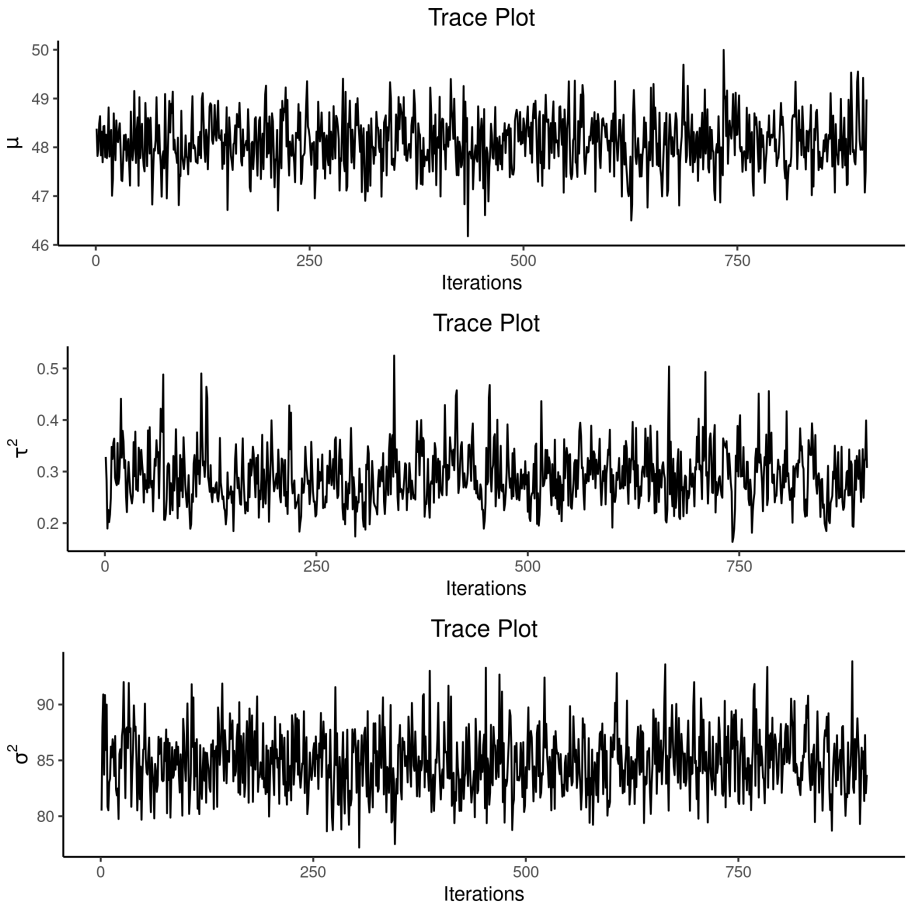

Figure 2: Trace Plots for μ, τ^2, σ^2.

D. Express the conditional posterior mean for θ_i in the following form:

$$E(\theta_i|y, \tau^2, \sigma^2, \mu) = \kappa_i \mu + (1 - \kappa_i)\bar{y}_i,$$

i.e. a convex combination of prior mean and data mean. Here κ_i is a *shrinkage coefficient* whose form you should express in terms of the model hyperparameters. Compute κ_i for each school in your MCMC algorithm.

From (c), we know that the mean in the full-conditional for θ_i is

$$\begin{aligned} E(\theta_i|y, \tau^2, \sigma^2, \mu) &= \left[\frac{N_i}{\sigma^2} + \frac{1}{\tau^2\sigma^2}\right]^{-1} \left(\frac{\sum_{j=1}^{N_i} y_{ij}}{\sigma^2} + \frac{\mu}{\tau^2\sigma^2}\right) \\ &= \left[\frac{N_i\tau^2 + 1}{\tau^2\sigma^2}\right]^{-1} \left(\frac{\sum_{j=1}^{N_i} y_{ij}}{\sigma^2} + \frac{\mu}{\tau^2\sigma^2}\right) \\ &= \frac{\tau^2 \sum_{j=1}^{N_i} y_{ij}}{N_i\tau^2 + 1} + \frac{\mu}{N_i\tau^2 + 1} \\ &= (1 - \kappa_i)\bar{y}_i + \kappa_i\mu, \end{aligned}$$

where $\kappa_i = \frac{1}{1+\tau^2 N_i}$.

E. Observe that an equivalent way to write your model involves the following decomposition:

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

where $\delta_i \sim N(0, \tau^2\sigma^2)$ and $e_{ij} \sim N(0, \sigma^2)$. (In the paper by Gelman that I've asked you to read, he writes it this way, where the school-level "offsets" are centered at zero, although he doesn't scale these offsets by σ the way I prefer to do.) To translate between the two parameterizations, just observe that in the previous version, $\theta_i = \mu + \delta_i$. Conditional on the "grand mean" μ, but *marginally* over both δ_i and e_{ij}, compute the following two covariances:

$$\mathbf{cov}(y_{i,j}, y_{i,k}),\ j \neq k$$
$$\mathbf{cov}(y_{i,j}, y_{i',k}),\ i \neq i'\ j \neq k$$

Does this make sense to you? Why or why not?

We can easily derive the desired expressions as follows:

$$\begin{aligned}
\text{cov}(y_{ij}, y_{ik}) &= \text{cov}(\mu + \delta_i + e_{ij}, \mu + \delta_i + e_{ik}) \\
&= E\left[(\delta_i + e_{ij})(\delta_i + e_{ik})\right] \\
&= E(\delta_i^2) \\
&= \tau^2\sigma^2, \\
\text{cov}(y_{ij}, y_{i'k}) &= \text{cov}(\mu + \delta_i + e_{ij}, \mu + \delta_{i'} + e_{i'k}) \\
&= E\left[(\delta_i + e_{ij})(\delta_{i'} + e_{i'k})\right] \\
&= 0,
\end{aligned}$$

where we assume independence between i and i' to obtain the penultimate equality. This makes sense to me since test scores for students in the same school are expected to be correlated while test scores for students at different schools should not be correlated.

F. Does the assumption that σ^2 is common to all schools look justified in light of the data?

It appears that this assumption is appropriate. We account for in-school variability with the scaling τ^2, while variability between schools is modeled with σ^2. This matches our intuition that the variability between schools and the variability within schools should be closely tied together.

A. Is the experimental medication effective at reducing blood pressure? Do the naive thing and perform a t-test for a difference of means, pooling all the data from treatment 1 into group 1, and all the data from treatment 2 into group 2. What does this t-test say about the difference between these two group means and the standard error for the difference? Why is the t-test (badly) wrong?

R output from a t-test for a difference in means can be found in Figure 3; from it, we can see that the difference between these two group means is 9.469 with a standard error of 1.004414. However, we should not trust this inference because we violate an assumption when using the t-test.

The t-test assumes that observations within groups are independent, which is clearly not the case here since we have repeated measurements of the same individuals. Perhaps if we had a single measurement for each individual within the groups, the t-test would be more valid. However, as our data stands, we should not perform a t-test for a difference in means.

```
        Welch Two Sample t-test

data:  group1 and group2
t = 9.4273, df = 391.66, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7.494212 11.443648
sample estimates:
mean of x mean of y
  142.455   132.986

> test$stderr
[1] 1.004414
```

Figure 3: R Output from a t-test with Pooled Observations.

B. Now, do something better but less than ideal. Calculate $\bar{y}_i$, the mean blood pressure measurement for each patient. Then, treat each person-level mean as if it were just a single data point, and conduct a different t-test for mean blood pressure between treatment 1 and treatment 2. What does this t-test say about the difference between these two group means and the standard error for the difference? Why is the standard error so much bigger, and why is this appropriate? Even so, why is this approach (subtly) wrong?

R output from a t-test for a difference in means using average blood pressure measurements for each patient can be found in Figure 4. In it, we can see an estimated difference between group means of 7.416 and a standard error of 4.511762.

The standard error here is bigger than it was in (A) because we are using less observations for our t-test, which more closely reflects reality. Although we have a total of 426 observations in this data set, we only have 20 individuals with repeated samples. Therefore, our sample size is actually much smaller than it first seems. For that reason, our standard error will be larger than it was when we told our t-test that we had a sample size of 426.

While a better approach to the t-test, this testing methodology is still not that great because we are not obtaining standard errors between individuals and between groups. To fully understand the data, we should account for differences within-subjects and within-groups. Here, we are just accounting for difference between groups.

```
        Welch Two Sample t-test

data:  group1 and group2
t = 1.6437, df = 17.09, p-value = 0.1185
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.099195 16.931139
sample estimates:
mean of x mean of y
 141.5435  134.1275

> testb$stderr
[1] 4.511762
```

Figure 4: R Output from a t-test with Mean Observations.

C. Now fit a two-level hierarchical model to this data, of the following form:

$$(y_{ij} \mid \theta_i, \sigma^2) \sim \mathbf{N}(\theta_i, \sigma^2)$$
$$(\theta_i \mid \tau^2, \sigma^2) \sim \mathbf{N}(\mu + \beta x_i, \tau^2\sigma^2)$$

where y_{ij} is blood pressure measurement j on person i, and x_i is a dummy (0/1) variable indicating whether a patient received treatment 2, the experimental medication. Apply what you learned on the previous problem about sampling, hyperparameters, etc, but account for the extra wrinkle here, i.e. the presence of the βx_i term that shifts the mean between the treatment and control groups.

Write out your model's complete conditional distributions, and fit it. Make a histogram of the posterior distribution for β, which represents the treatment effect here. In particular, what are the posterior mean and standard deviation of β? How do these compare to the estimates and standard errors from the approaches in (A) and (B)?

Before we fit the model via Gibbs sampler, we need to obtain the full-conditionals for $\mu, \tau^2, \sigma^2, \theta, \beta$:

$$\begin{aligned}
p(\mu|\cdot) &\propto \prod_{i=1}^{P} p(\theta_i|\mu, \tau^2\sigma^2) \cdot p(\mu) \\
&\propto \prod_{i=1}^{P} \exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i - \mu - \beta x_i)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\tau^2\sigma^2}\sum_{i=1}^{P}(-2\theta_i\mu + \mu^2 + 2\mu\beta x_i)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-2\left(\frac{\sum_{i=1}^{P}\theta_i - \beta\sum_{i=1}^{P}x_i}{\tau^2\sigma^2}\right)\mu + \mu^2\left(\frac{P}{\tau^2\sigma^2}\right)\right]\right\}, \\
p(\theta_i|\cdot) &\propto \prod_{j=1}^{N_i}\left[\exp\left\{-\frac{1}{2\sigma^2}(y_{ij} - \theta_i)^2\right\}\right]\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i - \mu - \beta x_i)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{N_i}(-2y_{ij}\theta_i + \theta_i^2)\right\}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i^2 - 2\mu\theta_i - 2\beta x_i\theta_i)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-2\left(\frac{\sum_{j=1}^{N_i}y_{ij}}{\sigma^2} + \frac{\mu + \beta x_i}{\tau^2\sigma^2}\right)\theta_i + \theta_i^2\left(\frac{N_i}{\sigma^2} + \frac{1}{\tau^2\sigma^2}\right)\right]\right\}
\end{aligned}$$

so, we see that

$$p(\mu|\cdot) \equiv N\left(\frac{\sum_{i=1}^{P}\theta_i - \beta\sum_{i=1}^{P}x_i}{P}, \frac{\tau^2\sigma^2}{P}\right)$$
$$p(\theta_i|\cdot) \equiv N\left(\left[\frac{N_i}{\sigma^2} + \frac{1}{\tau^2\sigma^2}\right]^{-1}\left(\frac{\sum_{j=1}^{N_i}y_{ij}}{\sigma^2} + \frac{\mu+\beta x_i}{\tau^2\sigma^2}\right), \left[\frac{N_i}{\sigma^2} + \frac{1}{\tau^2\sigma^2}\right]^{-1}\right)$$

Next, let's get those full-conditionals for σ^2 and τ^2:

$$\begin{aligned}
p(\sigma^2|\cdot) &\propto \prod_{i=1}^{P}\left[\prod_{j=1}^{N_i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y_{ij}-\theta_i)^2\right\}\cdot\frac{1}{\sqrt{2\pi\tau^2\sigma^2}}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i-\mu-\beta x_i)^2\right\}\right]\cdot\frac{1}{\sigma^2} \\
&\propto \sigma^{-2}\prod_{i=1}^{P}\left[\sigma^{-N_i}\exp\left\{-\frac{1}{\sigma^2}\sum_{j=1}^{N_i}\left(\frac{y_{ij}-\theta_i}{2}\right)^2\right\}\cdot\sigma^{-1}\exp\left\{-\frac{1}{\sigma^2}\left(\frac{(\theta_i-\mu-\beta x_i)^2}{2\tau^2}\right)\right\}\right] \\
&\propto (\sigma^2)^{-\left(\frac{\sum_{i=1}^{P}N_i+P}{2}\right)-1}\exp\left\{-\frac{1}{\sigma^2}\left[\sum_{i=1}^{P}\sum_{j=1}^{N_i}\left(\frac{(y_{ij}-\theta_i)^2}{2}\right)+\sum_{i=1}^{P}\left(\frac{(\theta_i-\mu-\beta x_i)^2}{2\tau^2}\right)\right]\right\}, \\
p(\tau^2|\cdot) &\propto \prod_{i=1}^{P}\left[\frac{1}{\sqrt{2\pi\tau^2\sigma^2}}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i-\mu-\beta x_i)^2\right\}\right]\cdot(\tau^2)^{-\frac{1}{2}-1}\exp\left\{-\frac{1}{2\tau^2}\right\} \\
&\propto (\tau^2)^{-\left(\frac{P+1}{2}\right)-1}\exp\left\{-\frac{1}{\tau^2}\left[\frac{\sum_{i=1}^{P}(\theta_i-\mu-\beta x_i)^2}{2\sigma^2}+\frac{1}{2}\right]\right\},
\end{aligned}$$

which means that we have the following full-conditionals for σ^2 and τ^2:

$$p(\sigma^2|\cdot) \equiv \text{IG}\left(\frac{\sum_{i=1}^{P}N_i+P}{2}, \sum_{i=1}^{P}\sum_{j=1}^{N_i}\left(\frac{(y_{ij}-\theta_i)^2}{2}\right)+\sum_{i=1}^{P}\left(\frac{(\theta_i-\mu-\beta x_i)^2}{2\tau^2}\right)\right),$$
$$p(\tau^2|\cdot) \equiv \text{IG}\left(\frac{P+1}{2}, \frac{\sum_{i=1}^{P}(\theta_i-\mu-\beta x_i)^2}{2\sigma^2}+\frac{1}{2}\right)$$

Lastly, we need to obtain the full-conditional for our β parameter, which we assume has a normal prior with mean μ_β and variance σ^2_β:

$$\begin{aligned}
p(\beta|\cdot) &\propto \prod_{i=1}^{P}\left[\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i-\mu-\beta x_i)^2\right\}\right]\cdot\exp\left\{-\frac{1}{2\sigma^2_\beta}(\beta-\mu_\beta)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\tau^2\sigma^2}\sum_{i=1}^{P}(-2\theta_i x_i\beta+2\mu x_i\beta+\beta^2x_i^2)\right\}\cdot\exp\left\{-\frac{1}{2\sigma^2_\beta}(\beta^2-2\mu_\beta\beta)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-2\left(\frac{\sum_{i=1}^{P}(\theta_i-\mu)x_i}{\tau^2\sigma^2}+\frac{\mu_\beta}{\sigma^2_\beta}\right)\beta+\beta^2\left(\frac{\sum_{i=1}^{P}x_i^2}{\tau^2\sigma^2}+\frac{1}{\sigma^2_\beta}\right)\right]\right\},
\end{aligned}$$

which is the kernel of a Normal distribution, so

$$p(\beta|\cdot) \equiv \text{Normal}\left(\left(\frac{\sum_{i=1}^{P} x_i^2}{\tau^2\sigma^2} + \frac{1}{\sigma_\beta^2}\right)^{-1}\left(\frac{\sum_{i=1}^{P}(\theta_i - \mu)x_i}{\tau^2\sigma^2} + \frac{\mu_\beta}{\sigma_\beta^2}\right), \left(\frac{\sum_{i=1}^{P} x_i^2}{\tau^2\sigma^2} + \frac{1}{\sigma_\beta^2}\right)^{-1}\right),$$

The posterior mean for β is -0.3901348 and the posterior standard deviation for β is 0.9825388 with $M = 1000$ iterations, seed 702, $\mu_\beta = 0$, $\sigma_\beta^2 = 1$, and a burn-in period of 100 iterations. A histogram of the posterior distribution for β can be found in Figure 5. Additionally, trace plots for the parameters in our hierarchical model can be found in Figure 6 for diagnostics.

In comparison to (A) and (B), where the difference in means was estimated to be roughly 8, our posterior estimate for β is -0.39; this means that if someone received the experimental treatment, then we expect to see a difference in blood pressure measurement of -0.39. This is significantly less than the inference we obtain from (A) and (B), which may be due to a larger number of subjects with systolic readings that are not representative of their group. Additionally, our standard error is slightly smaller than that in (A) and smaller than in (B).

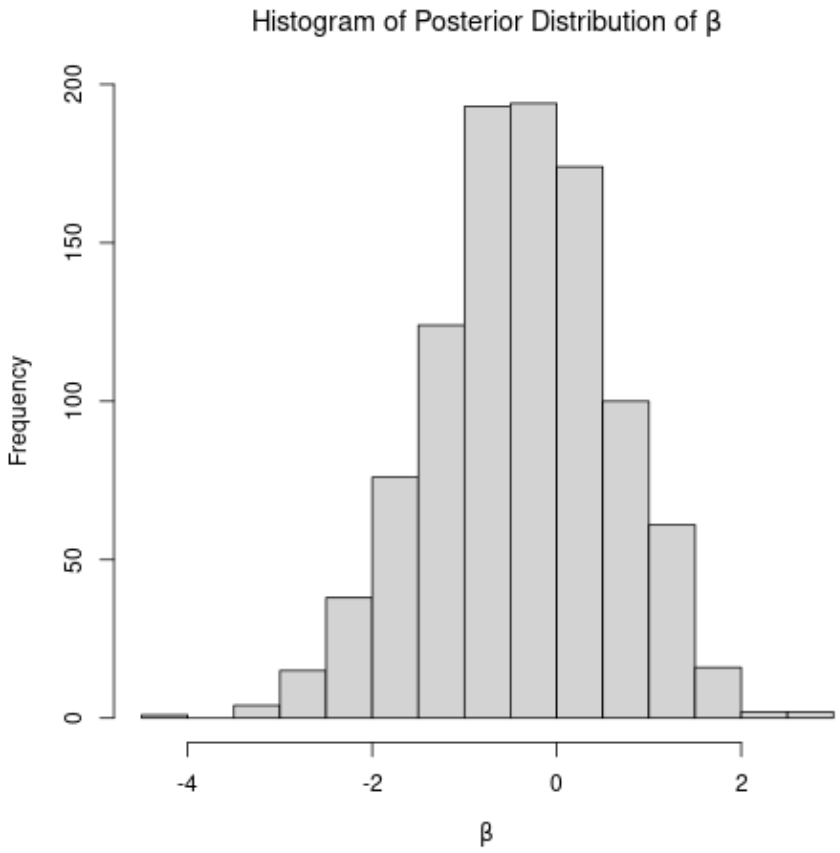

Figure 5: Histogram of Posterior for β.

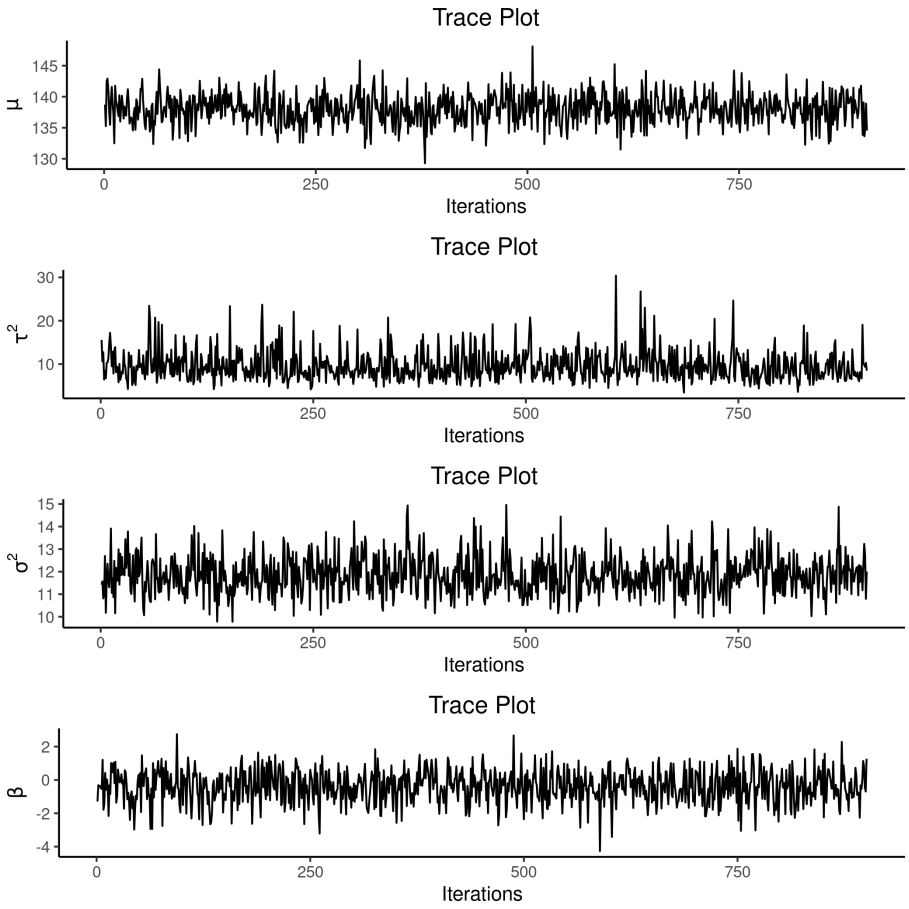

Figure 6: TRACES OF BLOOD!

D. Your two-level model assumes that, conditional on θ_i, the y_{ij} are independent. Written concisely: $(y_{ij} \perp y_{ik} \mid \theta_i)$ for $j \neq k$. There are many ways this assumption could break down. So check! Does this assumption look (approximately) sensible in light of the data? Provide evidence one way or another.

I believe this assumption to be insensible since blood pressure measurements appear to be correlated within individuals. For example, subjects 2 and 16 in Figure 7 appear to have some autocorrelation in their blood measurements. This indicates to me that, conditional on θ_i, the y_{ij} are not independent.

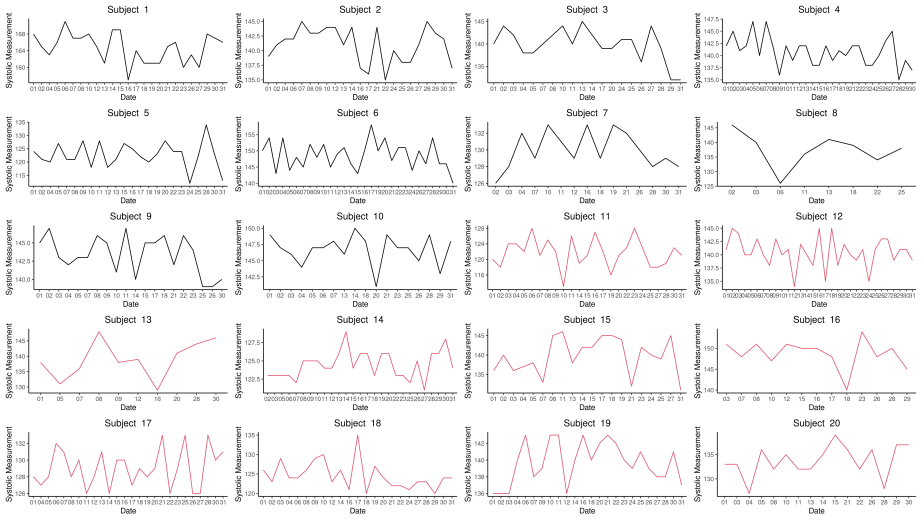

Figure 7: Subject-level blood measurements.