HW2

Sam Olson

Q1

Discuss whether you believe it would be better to view this problem as one involving a Bayesian analysis of a mixture model, or as one we should approach with a model having several levels of prior distributions. *Hint: Read Assignment 2 – Background carefully before developing your answer to this question.*

Answer

Note on Ch. 15

Choose the interpretation based on **what quantity is scientifically meaningful to estimate**. If your goal is inference about **specific units**, so that the group parameters θ_i represent persistent, interpretable characteristics and you could plausibly observe additional data generated under the same θ_i (e.g., the same hospital, school, or lake), then treat the hierarchy as a **multi-stage prior**,

$$f(y \mid \theta), \quad \pi_1(\theta \mid \lambda), \quad \pi_2(\lambda),$$

and focus on the unit-level posteriors $p(\theta_i \mid y)$ with partial pooling.

If instead your goal is inference about the **population distribution of effects across groups**, and the observed units are best viewed as exchangeable draws from that population so that the individual θ_i are not themselves of intrinsic interest, then treat the model as a **mixture**, integrate out θ to obtain

$$f(y \mid \lambda) = \int f(y \mid \theta)\, g(\theta \mid \lambda)\, d\theta,$$

and focus on $p(\lambda \mid y)$ or predictive distributions $p(\theta^* \mid y)$. In short: **unit-specific questions ⇒ infer θ_i; distributional/population questions ⇒ infer λ.**

Background Context

As described in the background material, biologists at the IDNR are concerned with assessing the **health (“condition”) of fish populations within lakes**, where condition reflects how heavy fish are for their length. Although individual fish vary, condition is primarily governed by **lake-level environmental factors** such as food availability and habitat quality, and is therefore interpreted as a **population- or lake-level characteristic**, not an individual trait. Ecologically, the objective is to identify whether some lakes exhibit poorer growth (e.g., “stunted” populations) and to understand how fish condition differs **across lakes**.

Let $Y_{i,j}$ denote the weight of fish j in lake i, and $x_{i,j}$ its length. The hierarchical weight–length model

$$Y_{i,j} = \mu_{i,j} + \sigma_i \mu_{i,j}^{\theta} \varepsilon_{i,j}, \qquad \mu_{i,j} = \beta_i x_{i,j}^{\alpha}, \qquad \varepsilon_{i,j} \overset{iid}{\sim} N(0,1),$$

includes a shared species parameter α and lake-specific parameters β_i (condition) and σ_i^2 (variability). These lake-specific effects are modeled hierarchically,

$$\beta_i \sim N(\lambda, \tau^2), \qquad \sigma_i^2 \sim IG(\xi_1, \xi_2),$$

allowing information to be shared across lakes. Importantly, the scientific goal is not to learn about any single fish or even a single lake in isolation, but rather to understand **how condition varies across the population of lakes** and to characterize the overall distribution of lake-level condition parameters.

Direct Answer

Because the scientific objective is **distributional**, namely to understand how fish condition varies across lakes and to characterize the population-level distribution of lake effects, this problem is most naturally viewed as a **Bayesian mixture model**. In this interpretation, the lake-specific parameters β_i are exchangeable realizations from a common distribution, and inference focuses primarily on the hyperparameters (λ, τ^2) and predictive distributions such as $p(\beta^* \mid y)$, which describe the typical level and variability of condition across lakes. Although the hierarchical formulation still yields posterior distributions for individual β_i, these are secondary summaries, while the primary inferential target is the **population distribution of lake health**, not any particular lake's parameter. Therefore, the mixture-model viewpoint provides the most appropriate conceptual framework for this problem.

Q2:

Based on your answer to question 1, what would you include in a summary of your inferences associated with the problem. Be specific about types of summary information (e.g., five number summary or table of quantiles) and/or graphs and plots (e.g., scatterplot of y versus x, plot of empirical distribution of f). Indicate how these inferences can be used to address the goals of the Iowa DNR.

Answer

Because the problem is most naturally interpreted under the **mixture (population-level) viewpoint**, inference should focus primarily on the **distribution of lake-specific condition parameters** rather than on any single lake in isolation. Thus, summaries should emphasize (i) the population distribution of condition across lakes, (ii) uncertainty in that distribution, and (iii) how individual lakes compare relative to the overall population.

Population-level summaries (primary targets)

Since the scientific question concerns *how condition varies across lakes*, the key inferential quantities are the hyperparameters governing the distribution of lake effects, (λ, τ^2), and the predictive distribution of a new lake's condition, $p(\beta^* \mid y)$.

I would report:

- Posterior means, standard deviations, and 95% credible intervals for (λ, τ^2)
- A five-number summary or table of posterior quantiles (5%, 25%, 50%, 75%, 95%) for β^*
- A density plot or histogram of draws from $p(\beta^* \mid y)$

These summarize: - the **typical lake condition** (location of λ), - the **between-lake variability** (magnitude of τ^2), - and the **expected range of condition** across Iowa lakes.

This directly addresses the DNR's goal of understanding overall lake health and variability statewide.

Lake-specific summaries (secondary but useful)

Although not the primary inferential target, the posterior distributions $p(\beta_i \mid y)$ provide descriptive comparisons among lakes.

For each lake, I would report:

- Posterior mean or median of β_i
- 95% credible interval
- A ranked table of posterior means
- Possibly standardized scores $(\beta_i - \lambda)/\tau$

Useful plots include:

- Caterpillar (forest) plot of β_i with credible intervals
- Boxplot or empirical distribution of $\{\beta_i\}$
- Posterior ranks or probability that β_i is among the lowest/highest

These allow the DNR to: - identify lakes with unusually poor condition, - flag candidates for intervention, - compare lakes relative to statewide norms.

Model-fit and diagnostic visualizations

To assess adequacy of the weight–length relationship and communicate results:

- Scatterplot of $Y_{i,j}$ vs $x_{i,j}$ with fitted curves $\hat{\mu}_{i,j} = \hat{\beta}_i x_{i,j}^{\hat{\alpha}}$ by lake
- Posterior predictive checks (simulated vs observed weights)
- Residual plots

These verify that the biological growth relationship is well captured.

Interpretation for the Iowa DNR

Together, these summaries allow the DNR to:

- quantify **typical fish condition statewide**,
- measure **how much lakes differ from each other**,
- determine whether variability is small (uniform health) or large (problematic heterogeneity),
- identify **specific lakes with unusually low condition** for management or restoration,
- predict condition in **unsampled lakes** using $p(\beta^* \mid y)$.

Thus, inference emphasizes the **distribution of lake health** (population-level mechanism) while still providing practical lake-by-lake comparisons for decision-making.

Q3:

A statistical issue in the use of the hierarchical model developed in Assignment 2 – Background is the fixed value of the power $\theta = 1.0$ used in the analysis. Describe how you would conduct an assessment of this modeling choice. In particular, there are two immediate alternatives to our choice. One is that a single value of θ should be adequate to reflect data behavior, but its value should be something other than $\theta = 1.0$. In this case, we might assign a prior (similarly to α), but this question is not about what we might choose for that prior, only how we might assess the output of analysis of the hierarchical model in files `LMBmcmc1.txt` and `LMBmcmc2.txt` and the actual data (in file `LMBdat_for601.txt`) to determine whether we are motivated to include a prior for θ.

The other possibility is that we might allow this power to have different values in different lakes, and make use of parameters $\{\theta_i : i = 1, \ldots, n\}$. These quantities would then need to be assigned a distribution, the parameters of which will be assigned a prior but, again, don't worry about what any of those distributions might be, only how we could determine whether there is evidence that a single value of θ, be that $\theta = 1$ or some other value, appears adequate or inadequate to reflect the behavior of the actual data.

Answer

The parameter θ controls how variability scales with the mean through

$$\mathrm{Var}(Y_{i,j} \mid \cdot) = \sigma_i^2 \mu_{i,j}^{2\theta},$$

so fixing $\theta = 1$ imposes a specific mean–variance relationship. To assess whether this choice is adequate, I used the posterior output from the existing hierarchical fit (with $\theta = 1$) together with the observed data to examine whether this scaling produces approximately homoscedastic, Gaussian residuals.

Diagnostic procedure

Using posterior means of $\mu_{i,j}$ and σ_i^2, I computed standardized residuals under $\theta = 1$,

$$r_{i,j}^{(1)} = \frac{y_{i,j} - \hat{\mu}_{i,j}}{\hat{\sigma}_i \hat{\mu}_{i,j}},$$

which should resemble iid $N(0, 1)$ if the scaling is correct. I examined:

- scatterplots of $r_{i,j}^{(1)}$ vs $\hat{\mu}_{i,j}$,
- scatterplots of $|r_{i,j}^{(1)}|$ vs $\hat{\mu}_{i,j}$,
- a normal Q–Q plot,
- and a log–variance regression

$$\log\left(\frac{(y_{i,j} - \hat{\mu}_{i,j})^2}{\hat{\sigma}_i^2}\right) = c + 2\theta \log(\hat{\mu}_{i,j}) + \varepsilon,$$

where the slope estimates 2θ.

To assess possible heterogeneity across lakes, I also fit this regression separately by lake to obtain lake-specific estimates $\hat{\theta}_i$.

Results

The standardized residual plots show a clear pattern: residual spread still increases with $\hat{\mu}_{i,j}$ even after dividing by $\hat{\mu}_{i,j}$. The residuals are not approximately homoscedastic, and the Q–Q plot shows strong heavy right tails, indicating lack of normality under $\theta = 1$.

The pooled log–variance regression gives an estimated slope of about 1.86, implying

$$\hat{\theta} \approx 0.93,$$

which differs noticeably from 1. This suggests that the assumed scaling μ^1 slightly over-corrects the variance.

The lake-specific diagnostics show substantial variation in $\hat{\theta}_i$ (roughly from about 0.5 to 2), indicating that the appropriate variance scaling may differ across lakes rather than being constant statewide.

Conclusions

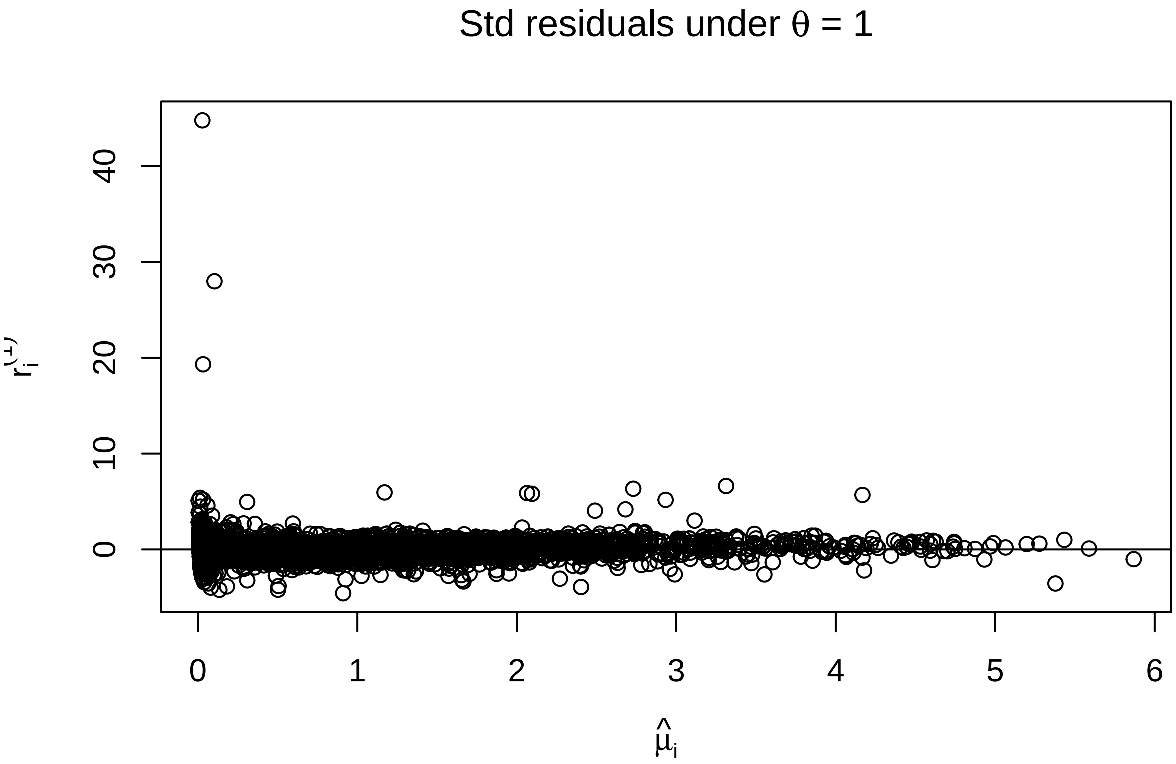
These diagnostics provide evidence that fixing $\theta = 1$ is not fully adequate:

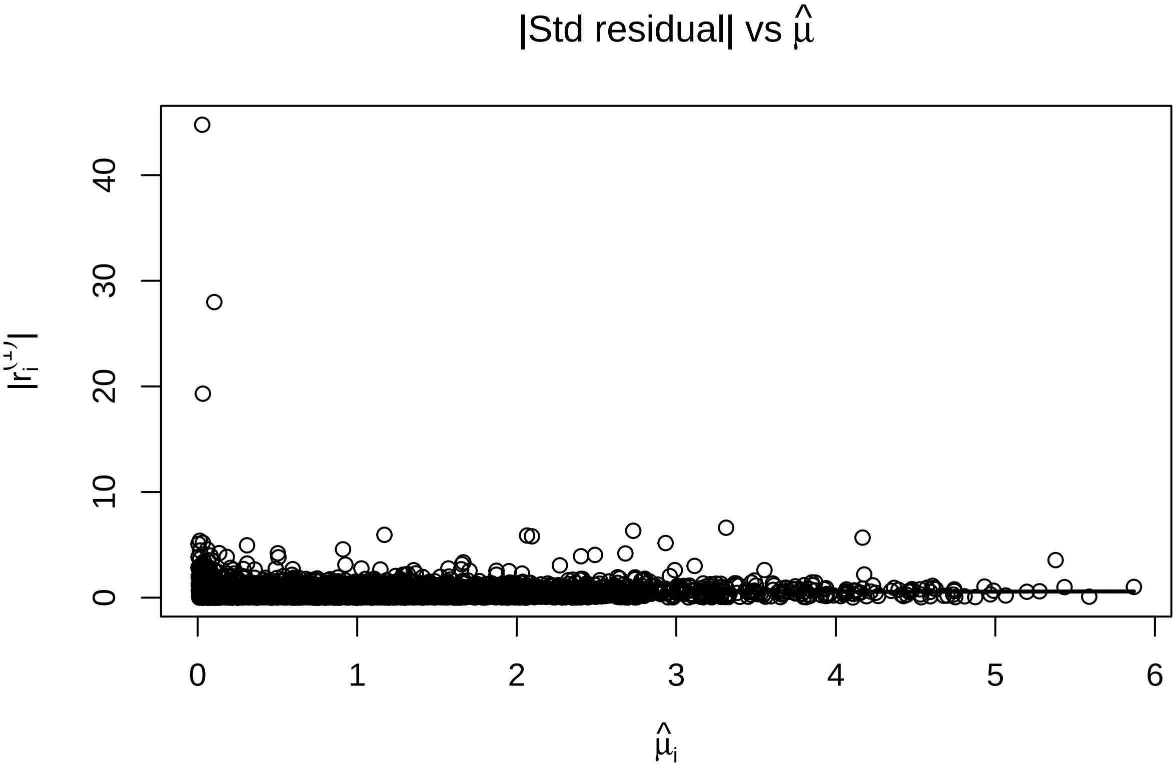
1. **Single common θ alternative:**
   Since the pooled estimate is clearly different from 1, there is motivation to treat θ as an unknown parameter and assign it a prior, estimating it jointly with the other parameters.

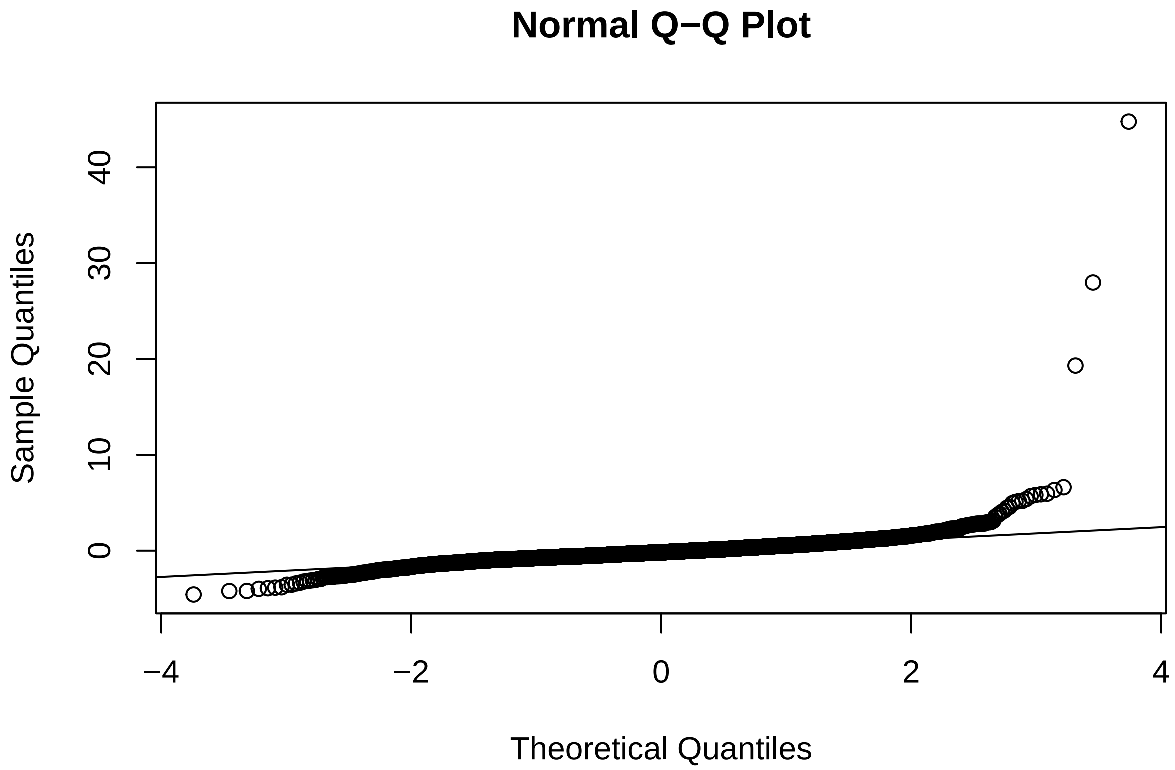
2. **Lake-specific θ_i alternative:**
   The variability in the lake-specific estimates suggests additional heterogeneity, so allowing $\{\theta_i\}$ to vary by lake may further improve fit if such flexibility is scientifically justified.

Overall, the data indicate that the fixed choice $\theta = 1$ does not fully capture the observed mean–variance relationship, and a hierarchical extension that estimates θ (globally or by lake) would likely better reflect the behavior of the data.

Appendix

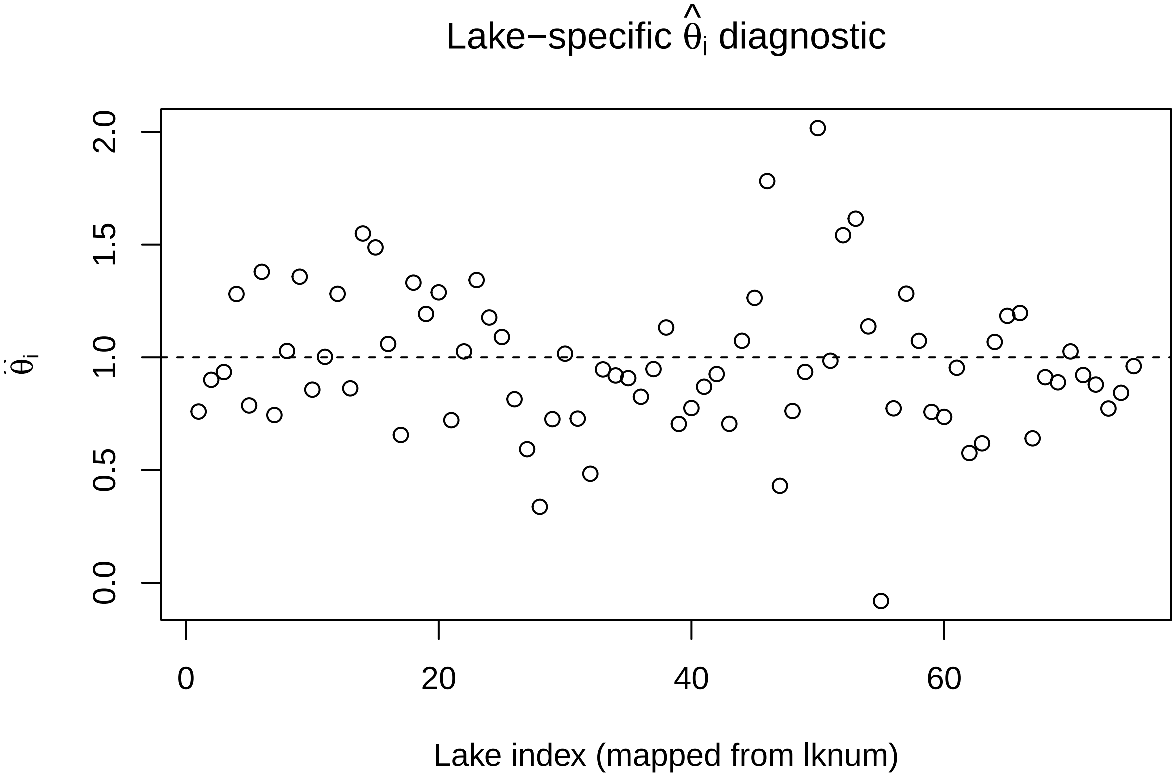

Std residuals under θ = 1
$r_i^{(\perp)}$
40
30
20
10
0
0
1
2
3
4
5
6
$\hat{\mu}_i$


|Std residual| vs $\hat{\mu}$
$|r_i^{(\perp)}|$
40
30
20
10
0
0
1
2
3
4
5
6
$\hat{\mu}_i$



```
##
## Call:
## lm(formula = y_star ~ x_star)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.6148  -1.0448   0.4899   1.4925   9.1900
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.08561    0.03563  -58.53   <2e-16 ***
## x_star       1.86010    0.02304   80.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.3 on 5452 degrees of freedom
## Multiple R-squared:  0.5446, Adjusted R-squared:  0.5445
## F-statistic:  6520 on 1 and 5452 DF,  p-value: < 2.2e-16
```

```
## [1] 0.9300478
```


Lake−specific $\hat{\theta}_i$ diagnostic
$\hat{\theta}_i$
0.0
0.5
1.0
1.5
2.0
0
20
40
60
Lake index (mapped from lknum)