Hierarchical Linear Models

Weight gain of pregnant women — individual, pooled and hierarchical regressions

Hedibert Freitas Lopes

Based on Example 5.7 of Gamerman, Lopes and Gonçalves (2026), *Markov Chain Monte Carlo*, 3nd ed.

Insper – Instituto de Ensino e Pesquisa
hedibert.org

These slides, the R code and all figures were prepared with AI assistance (Claude, Anthropic), from the prompt reproduced on the next slides.

The prompt that generated this deck (1/2)

The code below illustrates a simple linear and Gaussian hierarchical model. More information in the JPEG figure. Prepare slides latex/beamer that

1. *presents the data*
2. *fits Bayesian simple linear regression to each woman. Draw posterior densities for alphas (intercepts), betas (slopes) and sigmas (in the JPEG it shows a common sigma, but I want to change to one sigma per woman) to show whether they look similar.*
3. *Fit the pooled model, ignoring the distinction between the women. Compare the posterior densities of alpha, beta and sigma with the individual ones.*

The prompt that generated this deck (2/2)

4. *Fit the hierarchical model, but online the full conditional distributions. Add a hierarchy for sigma_i as well with common gamma prior Gamma(delta, eta), with Non informative hyperpriors for delta and eta.*
5. *Show the borrowing strength aspect of a hierarchical model*
6. *What would be the posterior predictive for a new woman (69th patient) for all weeks observed in the sample?*

. . . followed by the R data (`Y`, `x`, `grupo`) and a photo of Example 5.7.

Lesson: a clear, itemized, specific prompt (what to fit, what to plot, what to compare, which modification to the textbook model) is what makes the output usable.

Roadmap

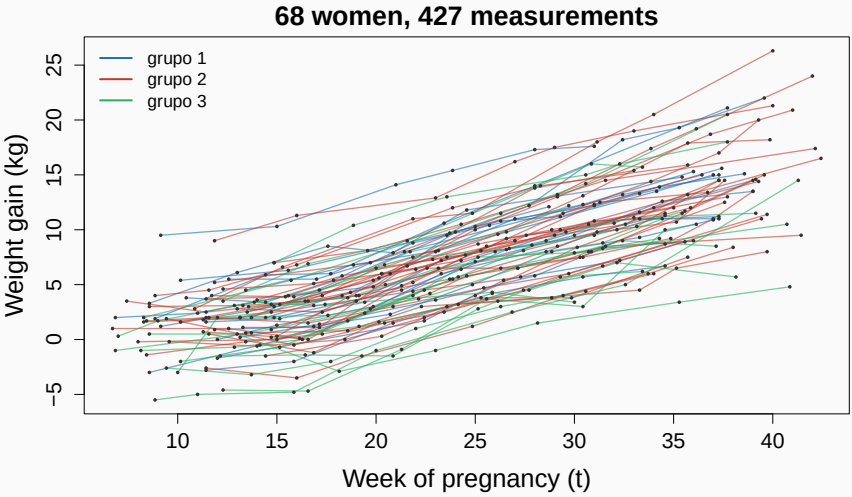
1. The data
2. Model 1: one regression per woman
3. Model 2: pooled regression
4. Model 3: hierarchical model
5. Borrowing strength
6. Prediction for a new woman
7. Summary

The data

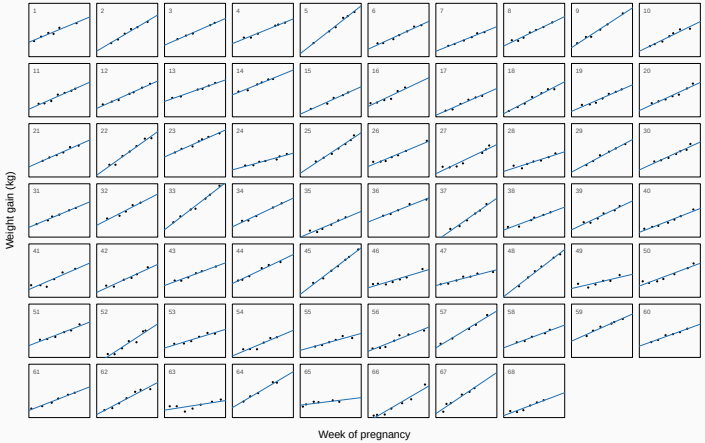
Weight gain during pregnancy

- Souza (1999): nutritional pattern of pregnant women followed at the *Instituto de Puericultura e Pediatria Martagão Gesteira* (UFRJ).
- $I = 68$ women, $n = \sum_i n_i = 427$ measurements:
  - y_{ij} = weight gain (kg) of woman i at visit j, ranging from -5.5 to 26.3 kg;
  - t_{ij} = week of pregnancy at visit j, ranging from 6.7 to 42.4.
- Number of visits per woman: $n_i = 5$ (8 women), $n_i = 6$ (33), $n_i = 7$ (27).
- Visits are **not** at common times: each woman has her own $t_{i1}, \ldots, t_{in_i}$.
- The R code also carries a group label (`grupo` $\in \{1, 2, 3\}$, with 19, 32 and 17 women); it is shown in the next plot but not used in the models.

All 68 trajectories



One woman per panel (with least-squares lines)



Within a woman, weight gain is close to linear in t; intercepts and slopes vary across women.

Model 1: one regression per woman

Individual Bayesian simple linear regressions

For each woman $i = 1, \ldots, 68$, *separately*:

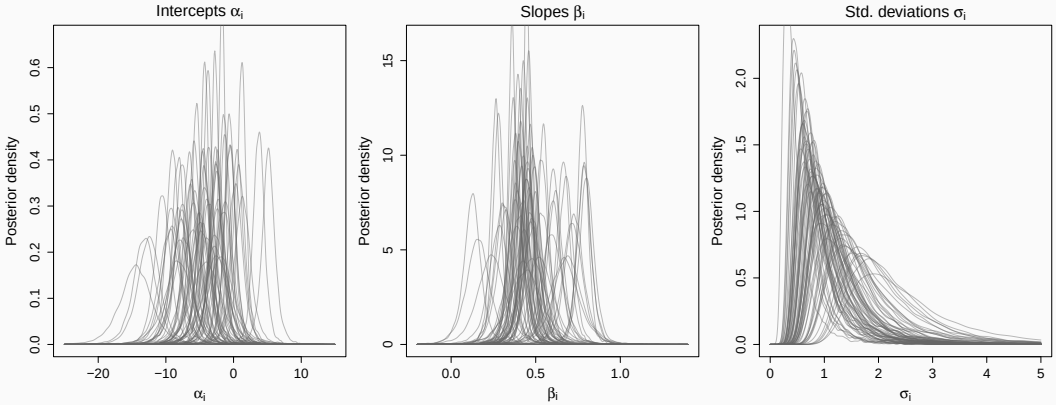
$$y_{ij} \mid \alpha_i, \beta_i, \sigma_i^2 \sim \mathcal{N}(\alpha_i + \beta_i t_{ij},\ \sigma_i^2), \qquad j = 1, \ldots, n_i, \qquad p(\alpha_i, \beta_i, \sigma_i^2) \propto \sigma_i^{-2}.$$

With $X_i = (\mathbf{1}, \boldsymbol{t}_i)$, $\hat{\boldsymbol{\theta}}_i = (X_i' X_i)^{-1} X_i' \boldsymbol{y}_i$ and $S_i = \|\boldsymbol{y}_i - X_i \hat{\boldsymbol{\theta}}_i\|^2$:

$$\sigma_i^2 \mid \boldsymbol{y}_i \sim \mathcal{IG}\left(\tfrac{n_i - 2}{2}, \tfrac{S_i}{2}\right), \qquad (\alpha_i, \beta_i)' \mid \sigma_i^2, \boldsymbol{y}_i \sim \mathcal{N}\left(\hat{\boldsymbol{\theta}}_i,\ \sigma_i^2 (X_i' X_i)^{-1}\right).$$

- Exact (i.i.d.) Monte Carlo: 10,000 draws per woman, no MCMC needed.
- Only $n_i - 2 \in \{3, 4, 5\}$ degrees of freedom $\Rightarrow$ heavy-tailed, wide posteriors, especially for σ_i.
- Time is not centred (as in the book), so α_i is the extrapolated gain at week 0 and is strongly negatively correlated with β_i a posteriori.

Posterior densities of α_i, β_i and σ_i (68 women)



Posterior medians: α_i from -14.4 to 5.2; β_i from 0.13 to 0.80 kg/week; σ_i from 0.39 to 2.31 kg. Do they look similar? **Similar enough to belong to one population, different enough not to be identical.**

Model 2: pooled regression

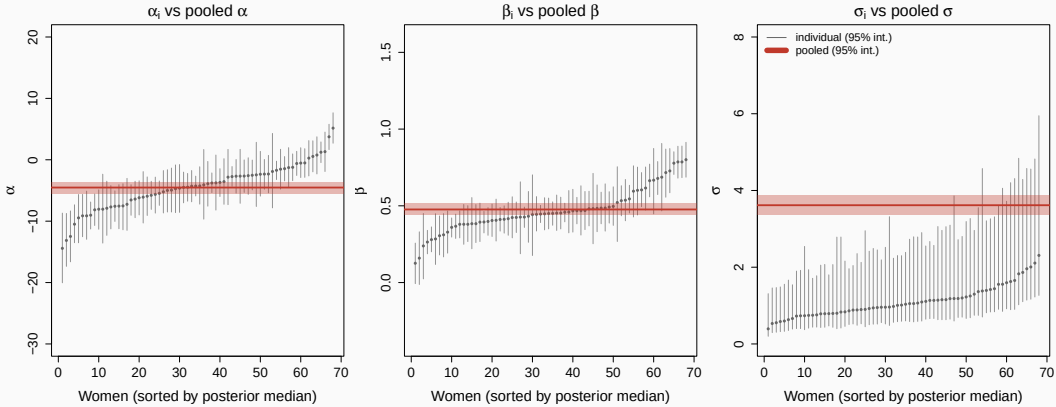
Pooled model: ignore who is who

$y_{ij} \mid \alpha, \beta, \sigma^2 \sim \mathcal{N}(\alpha + \beta t_{ij},\ \sigma^2)$, $\quad p(\alpha, \beta, \sigma^2) \propto \sigma^{-2}$, $\quad$ all $n = 427$ observations.

	posterior median (95% interval)	individual medians range
α	-4.52 (-5.49, -3.58)	-14.4 to 5.2
β	0.476 (0.439, 0.514)	0.13 to 0.80
σ	3.62 (3.38, 3.88)	0.39 to 2.31

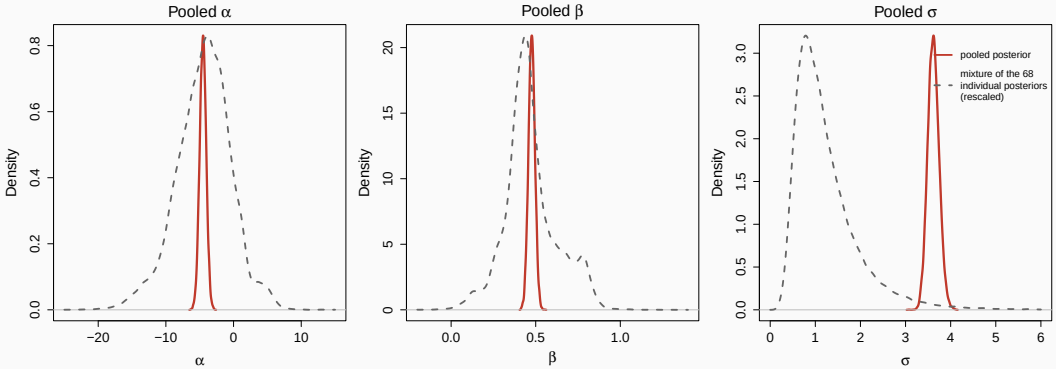
- α and β: very precise, sitting in the middle of the individual ones.
- σ: about 3.5 times the typical individual σ_i (median 1.02): the between-women variation in intercepts and slopes is dumped into the error term.

Individual 95% intervals vs pooled 95% interval



Many women are clearly incompatible with the pooled (α, β); and no woman has a σ_i as large as the pooled σ.

Pooled posterior vs the spread of the individual posteriors



Pooling gives false precision for (α, β) and a badly biased σ. Neither extreme (68 separate fits / 1 common fit) is satisfactory $\Rightarrow$ hierarchical model.

Model 3: hierarchical model

Hierarchical model with woman-specific variances

Write $\boldsymbol{\theta}_i = (\alpha_i, \beta_i)'$ and $\phi_i = \sigma_i^{-2}$. For $i = 1, \ldots, 68$, $j = 1, \ldots, n_i$:

Level 1: $y_{ij} \mid \boldsymbol{\theta}_i, \phi_i \sim \mathcal{N}(\alpha_i + \beta_i t_{ij},\ \phi_i^{-1})$

Level 2: $\alpha_i \mid \alpha, \tau_\alpha \sim \mathcal{N}(\alpha, \tau_\alpha^{-1}), \quad \beta_i \mid \beta, \tau_\beta \sim \mathcal{N}(\beta, \tau_\beta^{-1}), \quad \phi_i \mid \delta, \eta \sim \mathcal{G}(\delta, \eta)$

Level 3: $\alpha \sim \mathcal{N}(0, P_\alpha^{-1}),\ \beta \sim \mathcal{N}(0, P_\beta^{-1}), \quad \tau_\alpha, \tau_\beta \sim \mathcal{G}(a, b), \quad \delta \sim \mathcal{G}(c, d),\ \eta \sim \mathcal{G}(c, d)$

with $P_\alpha = P_\beta = 0.001$ and $a = b = c = d = 0.001$ (vague). $\mathcal{G}(\cdot, \cdot)$ = shape, rate.

- Book version (Example 5.7): a single σ^2, i.e. the limit $\delta \to \infty$ with η/δ fixed.
- Here the σ_i's are exchangeable: $E(\phi_i) = \delta/\eta$, $CV(\phi_i) = 1/\sqrt{\delta}$.
- δ measures how *similar* the women's variances are; η/δ sets their *level*.

Full conditionals (1/2): regression coefficients

Let $X_i = (\mathbf{1}, \boldsymbol{t}_i)$, $\boldsymbol{\mu} = (\alpha, \beta)'$, $D^{-1} = \mathrm{diag}(\tau_\alpha, \tau_\beta)$.

- **Woman-specific coefficients** (bivariate normal, drawn jointly):

$$\boldsymbol{\theta}_i \mid \cdot \;\sim\; \mathcal{N}(\boldsymbol{m}_i, V_i), \quad V_i = \left(\phi_i X_i' X_i + D^{-1}\right)^{-1}, \quad \boldsymbol{m}_i = V_i \left(\phi_i X_i' \boldsymbol{y}_i + D^{-1}\boldsymbol{\mu}\right).$$

- **Population means:**

$$\alpha \mid \cdot \sim \mathcal{N}\left(\frac{\tau_\alpha \sum_i \alpha_i}{I\tau_\alpha + P_\alpha},\; \frac{1}{I\tau_\alpha + P_\alpha}\right), \qquad \beta \mid \cdot \sim \mathcal{N}\left(\frac{\tau_\beta \sum_i \beta_i}{I\tau_\beta + P_\beta},\; \frac{1}{I\tau_\beta + P_\beta}\right).$$

- **Population precisions:**

$$\tau_\alpha \mid \cdot \sim \mathcal{G}\left(a + \tfrac{I}{2},\; b + \tfrac{1}{2}\textstyle\sum_i (\alpha_i - \alpha)^2\right), \qquad \tau_\beta \mid \cdot \sim \mathcal{G}\left(a + \tfrac{I}{2},\; b + \tfrac{1}{2}\textstyle\sum_i (\beta_i - \beta)^2\right).$$

Full conditionals (2/2): the variance hierarchy

- **Woman-specific precisions:**

$$\phi_i \mid \cdot \sim \mathcal{G}\Big(\delta + \tfrac{n_i}{2},\ \eta + \tfrac{1}{2}\textstyle\sum_j (y_{ij} - \alpha_i - \beta_i t_{ij})^2\Big).$$

- **Rate:** $\eta \mid \delta, \boldsymbol{\phi} \sim \mathcal{G}(c + I\delta,\ d + \sum_i \phi_i)$.
- **Shape:** not a standard distribution,

$$p(\delta \mid \eta, \boldsymbol{\phi}) \ \propto\ \delta^{c-1} e^{-d\delta}\ \frac{\eta^{I\delta}}{\Gamma(\delta)^I}\ \Big(\prod_i \phi_i\Big)^{\delta}.$$

  In practice we draw (δ, η) *jointly*: integrate η out,

$$p(\delta \mid \boldsymbol{\phi}) \ \propto\ \delta^{c-1} e^{-d\delta}\ \frac{\Gamma(c + I\delta)}{\Gamma(\delta)^I\,(d + \sum_i \phi_i)^{c+I\delta}}\ \Big(\prod_i \phi_i\Big)^{\delta},$$

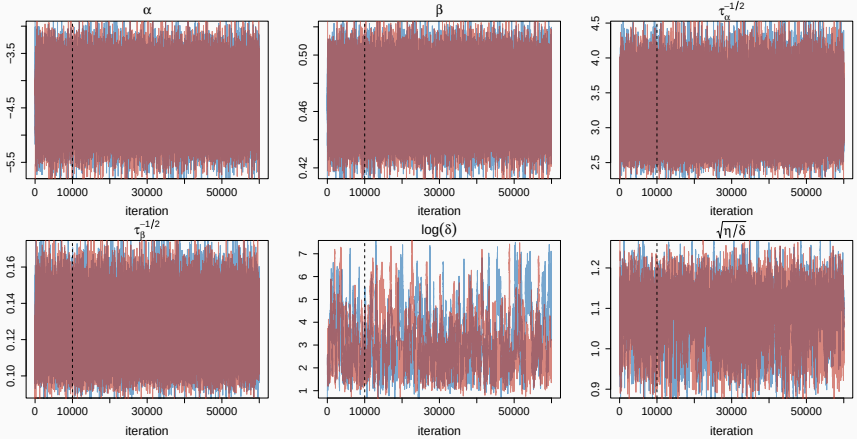
  update $\log\delta$ by random-walk Metropolis (acceptance $\approx 33\%$), then draw $\eta \mid \delta, \boldsymbol{\phi}$.
- Metropolis-within-Gibbs; the collapsed step avoids the strong (δ, η) posterior correlation.

The sampler in R (core of one sweep)

```
# (alpha_i, beta_i) jointly, vectorised over the 68 women (2x2 algebra by hand)
P11 <- phi*S11 + tau_a; P12 <- phi*S12; P22 <- phi*S22 + tau_b; det <- P11*P22 - P12^2
V11 <- P22/det; V12 <- -P12/det; V22 <- P11/det
g1 <- phi*r1 + tau_a*alpha; g2 <- phi*r2 + tau_b*beta
m1 <- V11*g1 + V12*g2; m2 <- V12*g1 + V22*g2
L11 <- sqrt(V11); L21 <- V12/L11; L22 <- sqrt(V22 - L21^2)
z1 <- rnorm(I); z2 <- rnorm(I); ai <- m1 + L11*z1; bi <- m2 + L21*z1 + L22*z2
# population level
pa <- I*tau_a + P; alpha <- rnorm(1, tau_a*sum(ai)/pa, 1/sqrt(pa))
pb <- I*tau_b + P; beta <- rnorm(1, tau_b*sum(bi)/pb, 1/sqrt(pb))
tau_a <- rgamma(1, a0 + I/2, b0 + sum((ai-alpha)^2)/2)
tau_b <- rgamma(1, a0 + I/2, b0 + sum((bi-beta)^2)/2)
# variance hierarchy
SSE <- syy - 2*ai*r1 - 2*bi*r2 + ai^2*S11 + 2*ai*bi*S12 + bi^2*S22
phi <- rgamma(I, delta + n/2, eta + SSE/2)
slp <- sum(log(phi)); sphi <- sum(phi); ldp <- log(delta) + 0.6*rnorm(1)
if (log(runif(1)) < ldelta_target(ldp,sphi,slp) - ldelta_target(log(delta),sphi,slp)) delta <- exp(ldp)
eta <- rgamma(1, c0 + I*delta, d0 + sphi)
```

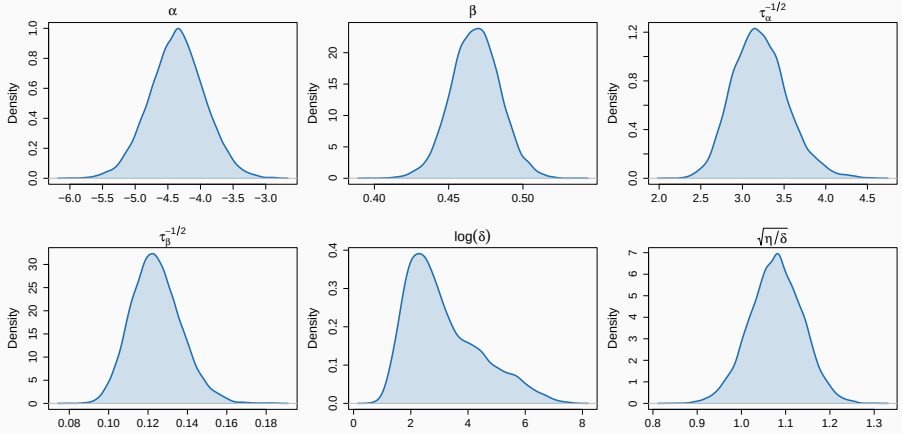
`S11, S12, S22, r1, r2, syy`: per-woman sufficient statistics $n_i, \sum_j t_{ij}, \sum_j t_{ij}^2, \sum_j y_{ij}, \sum_j t_{ij} y_{ij}, \sum_j y_{ij}^2$.

Convergence: two chains from dispersed starting values



60,000 iterations per chain, burn-in 10,000 (dashed line), thinning 10; 10,000 draws kept in total.
Gelman–Rubin $\hat{R} \leq 1.007$ for all six quantities.

Population-level posteriors



Population-level summaries

	median (95% interval)
α	-4.35 (-5.18, -3.53)
β	0.468 (0.436, 0.501)
$\tau_\alpha^{-1/2}$	3.19 (2.64, 3.93)
$\tau_\beta^{-1/2}$	0.123 (0.102, 0.151)
δ	16.52 (3.93, 468.79)
$1/\sqrt{\delta} = CV(\phi_i)$	0.25 (0.05, 0.50)
$\sqrt{\eta/\delta}$	1.08 (0.96, 1.19)

- α, β: close to the pooled values, *slightly* wider intervals.
- Real heterogeneity across women: sd of intercepts $\approx$ 3.2 kg, sd of slopes $\approx$ 0.12 kg/week.
- δ is large and its right tail is driven by the prior (the likelihood is flat as $\delta \to \infty$): the data say the σ_i are very similar, $CV(\phi_i)$ at most $\approx$ 0.5.
- So the book's common σ ($\approx \sqrt{\eta/\delta} \approx$ 1.1 kg) is a sensible simplification.

Borrowing strength

Where does "borrowing strength" come from?

Conditional on the hyperparameters, the full conditional mean of $\boldsymbol{\theta}_i$ is a **precision-weighted average**:

$$\boldsymbol{m}_i = \big(\underbrace{\phi_i X_i' X_i}_{\text{woman } i} + \underbrace{D^{-1}}_{\text{population}}\big)^{-1}\big(\phi_i X_i' X_i\, \hat{\boldsymbol{\theta}}_i + D^{-1}\boldsymbol{\mu}\big).$$

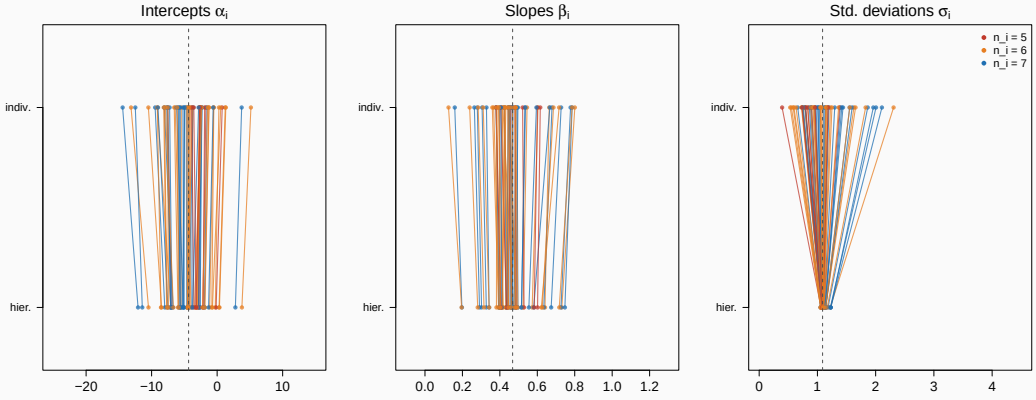
- Few or noisy observations (small $\phi_i X_i' X_i$) $\Rightarrow$ $\boldsymbol{\theta}_i$ pulled towards $\boldsymbol{\mu}$.
- Extreme $\hat{\boldsymbol{\theta}}_i \Rightarrow$ pulled back towards the others.

For the variances,

$$E(\phi_i \mid \cdot) = \frac{\delta + n_i/2}{\eta + S_i/2}: \qquad \text{the population contributes the equivalent of } 2\delta \text{ observations.}$$

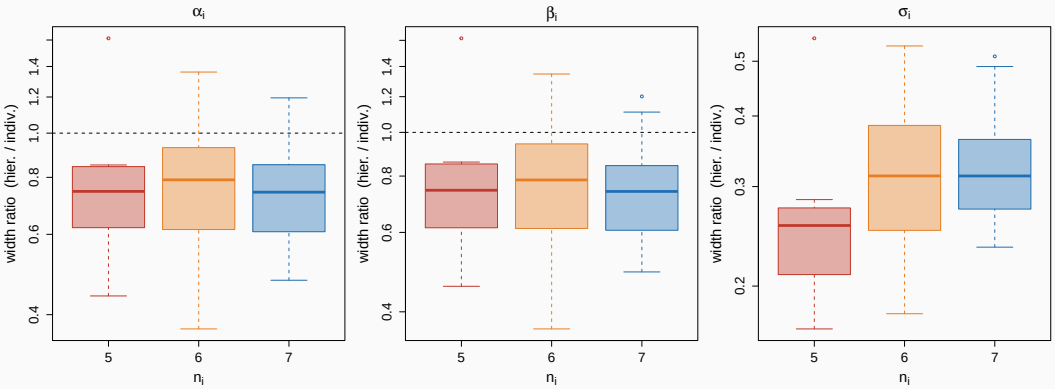
With $\delta \approx 16$ (≈ 33 pseudo-observations) vs $n_i \in \{5, 6, 7\}$: very strong pooling of the σ_i.

Shrinkage of posterior medians: individual → hierarchical



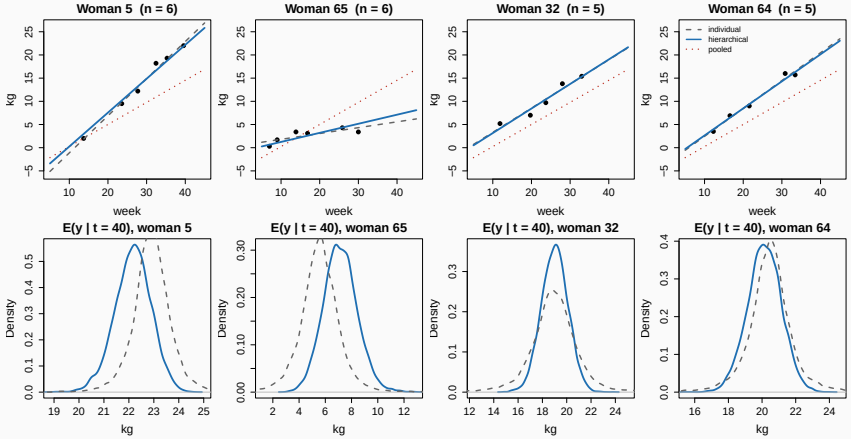
Dashed line: population centre. Standard deviation across women of the posterior medians: α_i 3.72 → 2.98; β_i 0.142 → 0.115; σ_i 0.40 → 0.04.

Precision gained: width of 95% intervals, hierarchical / individual



Median ratio: α_i 0.76, β_i 0.77, σ_i 0.30. Intervals for σ_i shrink for every woman (to about one third); for α_i, β_i most women gain, although a few with very well-determined lines get slightly wider intervals (their extreme values are pulled in, at a cost).

Four women: the extreme slopes and two with only five visits



Top: fitted lines (posterior medians). Bottom: posterior of the expected gain at week 40 — hierarchical (solid) is pulled towards the population and is narrower than individual (dashed).

Prediction for a new woman

Posterior predictive for woman 69

For a new woman observed at weeks $t = 7, \ldots, 42$ (the observed range):

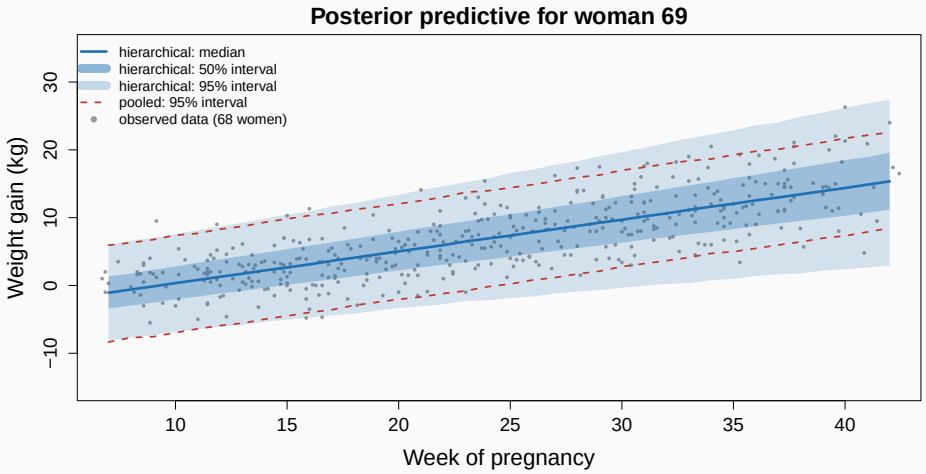
$$p(y_{69}(t) \mid \boldsymbol{y}) = \int \mathcal{N}\big(y_{69}(t); \alpha_{69} + \beta_{69} t,\ \phi_{69}^{-1}\big)\, p(\alpha_{69}, \beta_{69}, \phi_{69} \mid \boldsymbol{\psi})\, p(\boldsymbol{\psi} \mid \boldsymbol{y})\, d(\cdot),$$

$\boldsymbol{\psi} = (\alpha, \beta, \tau_\alpha, \tau_\beta, \delta, \eta)$. Composition sampling, for each MCMC draw $\boldsymbol{\psi}^{(m)}$:

1. $\alpha_{69}^{(m)} \sim \mathcal{N}(\alpha^{(m)}, 1/\tau_\alpha^{(m)})$, $\beta_{69}^{(m)} \sim \mathcal{N}(\beta^{(m)}, 1/\tau_\beta^{(m)})$, $\phi_{69}^{(m)} \sim \mathcal{G}(\delta^{(m)}, \eta^{(m)})$;
2. $y_{69}^{(m)}(t) \sim \mathcal{N}\big(\alpha_{69}^{(m)} + \beta_{69}^{(m)} t,\ 1/\phi_{69}^{(m)}\big)$ for every t.

- Uses *only* population-level information: woman 69 has no data yet.
- Three sources of uncertainty: $\boldsymbol{\psi}$, her own $(\alpha_{69}, \beta_{69}, \sigma_{69})$, and measurement noise.
- Implied σ_{69}: 1.10 (0.80, 1.53) kg.

Predictive bands for woman 69



The hierarchical band fans out with t (slope heterogeneity); the pooled band has constant width.

Predictive summaries at selected weeks (kg)

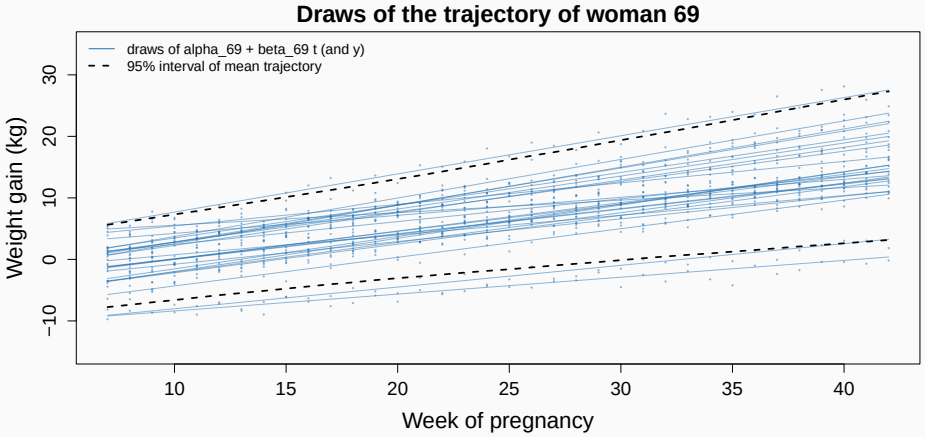
	hierarchical predictive					pooled
week	med.	25%	75%	2.5%	97.5%	95% int.
10	0.4	-2.0	2.8	-6.9	7.6	(-7.0, 7.4)
15	2.6	0.1	5.4	-5.0	10.5	(-4.5, 9.8)
20	5.0	2.2	7.9	-3.3	13.4	(-2.1, 12.0)
25	7.3	4.2	10.5	-1.9	16.6	(0.2, 14.4)
30	9.7	6.3	13.2	-0.3	19.7	(2.8, 17.0)
35	12.0	8.3	15.8	1.0	22.9	(5.0, 19.3)
40	14.4	10.3	18.5	2.4	26.2	(7.3, 21.7)

Share of the 427 observed weights inside the 95% band of a new woman:

weeks	hier.	pooled
all	98.4%	94.1%
$t \leq 20$	96.9%	96.2%
$20 < t \leq 30$	99.3%	94.9%
$t > 30$	99.2%	90.7%

Bands agree early; the pooled band is too narrow late in pregnancy. The hierarchical band is somewhat conservative.

What a “new woman” looks like: 25 predictive draws



Each draw is a plausible woman: her own line plus her own noise level.

Summary

Take-aways

1. **No pooling** (68 regressions): honest about heterogeneity, but $n_i \leq 7$ gives very wide posteriors, especially for σ_i.
2. **Complete pooling**: precise but wrong — σ absorbs the between-women variation (≈ 3.6 vs ≈ 1 kg).
3. **Hierarchical model**: each woman's estimate is a compromise between her own data and the population; the compromise is *learned from the data* (via $\tau_\alpha, \tau_\beta, \delta$).
4. All full conditionals are normal or gamma, except δ (one Metropolis step).
5. The variance hierarchy reveals that the σ_i are very similar: the data support the book's common σ, now as a *conclusion* rather than an assumption.
6. Predicting a new woman requires the hierarchy: the pooled model has no notion of "a new woman".

References

- Gamerman, D. , Lopes, H. F. and Gonçalves, F. B. (2026). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 3rd ed. Chapman & Hall/CRC. Example 5.7, pp. 157–161.
- Souza, A. D. P. (1999) *Approximate Methods in Bayesian Dynamic Hierarchical Models*. PhD thesis, COPPE-UFRJ. Unpublished Ph.D Thesis (in Portuguese).
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–472.

Files: `souza_hierarchical.R` reproduces every number and figure in these slides (it writes `results.tex`, `predtable.tex` and `figs/*.pdf`).