03/10/22

1. HW4 Solution ↑ w/ codes.

2. Student Evaluation Survey

† Random Effects (Hoff 8.2.1)

- Let Y_{ij} $j-$th observation from group i, $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$.
- $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})$ is a random sample of size m_i from group j.

  ⋆⋆ reasonable to assume $y_{ij} \mid \phi_i \overset{iid}{\sim} p(y \mid \phi_i)$.

  ⋆⋆ i.e., y_{ij} within group i are exchangeable, and y_{ij} and $y_{i'k}$, $i \neq i'$ are independent given ϕ_i's.
- the groups are samples from some larger population of groups

  $\Rightarrow$ reasonable to assume $\phi_i \mid \psi \overset{iid}{\sim} p(\phi \mid \psi)$.

  ⋆⋆ then place a prior for ψ, $\psi \sim p(\psi)$.

† Random Effects (Hoff 8.4): Match scores in U.S. public schools

- 2002 Educational Longitudinal Study (ELS), a survey of students from a large sample of schools across the U.S.

- Y_{ij}: math score of student j from school i.

  ⋆⋆ y_{ij} within group i are conditionally iid given some group specific parameters.

  $j = 1, \ldots, m_i$

  $\Rightarrow$ Let $y_{ij} \mid \theta_i, \sigma^2 \overset{iid}{\sim} \mathrm{N}(\theta_i, \sigma_i^2)$: within-group model

- The groups themselves are sample from some larger population of groups.

  ⋆⋆ school-level means θ_i are viewed as random effects arising from a normal population

  $\Rightarrow$ Let $\theta_i \mid \mu, \tau^2 \overset{iid}{\sim} \mathrm{N}(\mu, \tau^2)$: between-group model

† Random Effects (Hoff 8.4): Match scores in U.S. public schools

- Specify priors for unknown parameters $\underline{\sigma^2}$, $\underline{\mu}$ and $\underline{\tau^2}$;

$$\sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma), \quad \tau^2 \sim \mathrm{IG}(a_\tau, b_\tau), \quad \mu \sim \mathrm{N}(\mu_0, v_0^2).$$

- Inferential question of interest: Estimation of θ_i, or difference $\underbrace{\theta_i - \theta_{i'}}$.

$\widehat{(\theta_i - \theta_{i'})} = \frac{1}{B} \sum_{b=1}^{B} \left(\theta_i^{(b)} - \theta_{i'}^{(b)} \right)$, $\quad i \neq i'$

$\delta_{ii'} = \theta_i - \theta_{i'}$

$\underline{\delta_{ii'}^{(b)}} = \theta_i^{(b)} - \theta_{i'}^{(b)}$

† Random Effects (Hoff 8.4, BDA §15)

$\theta_i = \underline{\mu} + \gamma_i$

grand mean — μ; deviation from μ for group i — γ_i

- We may rewrite the model

fixed — random effects $\Rightarrow$ mixed effect model.

$$y_{ij} \mid \mu, \gamma_i \overset{indep}{\sim} \text{N}(\mu + \gamma_i, \sigma^2)$$

$$\mu \sim \text{N}(\mu_0, v_0^2), \quad \gamma_i \mid \tau^2 \overset{iid}{\sim} \text{N}(0, \tau^2), \quad (\sigma^2, \tau^2) \sim p(\cdot).$$

- school-level mean: overall mean μ plus some normal random effect $\gamma_i \Rightarrow$ mixed effects model.
- Note;

$$\text{Cov}(y_{ij}, y_{ij'}) = \tau^2, \text{ and } \quad \text{Cov}(y_{ij}, y_{i',j'}) = 0.$$

$\Rightarrow$ students within schools are exchangeable

$\Rightarrow$ student achievements across different schools are independent given the school effect

$j, k \in \{1, \dots, m_i\}$

$j \neq k$

Given μ, σ^2, τ^2

$$Y_{ij} = \mu + \gamma_i + \varepsilon_{ij}$$

$$Y_{ik} = \mu + \gamma_i + \varepsilon_{ik}$$

indep $\left[\begin{array}{l} \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2) \\ \gamma_i \overset{iid}{\sim} N(0, \tau^2) \end{array} \right.$

i : school

j : student

$$\mathrm{Cov}(Y_{ij}, Y_{ik}) = \mathrm{Cov}(\mu + \gamma_i + \varepsilon_{ij}, \mu + \gamma_i + \varepsilon_{ik})$$

$$= \mathrm{Cov}(\gamma_i, \gamma_i) + \mathrm{Cov}(\gamma_i, \varepsilon_{ik}) + \mathrm{Cov}(\varepsilon_{ij}, \gamma_i)$$

$$+ \mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ik})$$

$$= \tau^2 + 0 + 0 + 0 > 0$$

$$\mathrm{Var}(Y_{ij}) = \sigma^2 + \tau^2$$

$$\mathrm{Corr}(Y_{ij}, Y_{ik}) = \frac{\tau^2}{\sigma^2 + \tau^2} > 0$$

BDA

~~intra~~ class correlation

$i \neq i'$

$$\mathrm{Cov}(Y_{ij}, Y_{i'k}) = \underset{0}{\mathrm{Cov}(\gamma_i, \gamma_{i'})} + \cdots\cdots$$

$$= 0$$

† Varying-Coefficients Model (BDA §15)

- Souza (1999) considers a number of hierarchical models to describe the nutritional pattern of pregnant women. One of the models adopted was a hierarchical regression model where

$$\begin{aligned} y_{i,j} &\sim \mathrm{N}(\alpha_i + \beta_i t_{i,j}, \sigma^2), \\ (\alpha_i, \beta_i)' \mid \alpha, \beta &\sim \mathrm{N}_2((\alpha, \beta)', diag(\tau_\alpha^2, \tau_\beta^2)), \\ (\alpha, \beta)' &\sim \mathrm{N}_2((0,0)', diag(P_\alpha^2, P_\beta^2)). \end{aligned}$$

  Here $y_{i,j}$ and $t_{i,j}$ are the jth weight measurement and visit time of the ith woman with $j = 1 : n_i$ and $i = 1 : I$ for $I = 68$ pregnant women. Here $n = \sum_{i=1}^{I} n_i = 415$. For unknown scale parameters, we assume a priori independence and place inverse Gamma priors,

$$\sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma), \quad \tau_\alpha^2 \sim \mathrm{IG}(a_\alpha, b_\alpha), \text{ and } \tau_\beta^2 \sim \mathrm{IG}(a_\beta, b_\beta).$$

  Hyperparameters, $a_\sigma, b_\sigma, a_\alpha, b_\alpha, a_\beta, b_\beta$ P_α^2 P_β^2 are fixed.

† Model Choice - CR 7

- Suppose several models are in competition,

$$\mathcal{M}_i : x \sim f_i(x \mid \theta_i), \quad \theta_i \in \Theta_i, i \in I = \{1, \ldots, p\}.$$

- Model choice can be considered a special case of testing.

- The problem is not so simple since *while no model is true, several models may be appropriate*.

• **Example 7.1.1** Consider the data set relating the monthly unemployment rate with the monthly number of accidents in Michigan from 1978 to 1987. We may consider the following two models for the number of accidents N in a given month,

$$\begin{aligned} \mathcal{M}_1 : N &\sim \text{Poi}(\lambda), \lambda > 0. \\ \mathcal{M}_2 : N &\sim \text{NB}(m, p), m > 0 \text{ and } p \in [0, 1]. \end{aligned}$$

• **Example 7.1.2:** The dataset consists in 82 observations of galaxy velocities. For astrophysical reasons, the distribution of this dataset can be represented as a mixture of normal distributions whose number of components k is unknown. Reversable jump MCMC

$$\mathcal{M}_i : y_j \overset{iid}{\sim} \sum_{\ell=1}^{i} p_{\ell i} \mathrm{N}(\mu_{\ell i}, \sigma^2_{\ell i}), \quad j = 1, \ldots, 82.$$

Here i varies between 1 and some arbitrary upper bound.

⋆⋆ Note that a k component model is a submodel of a $(k+p)$ component mixture by letting the the p remaining components have weights 0.

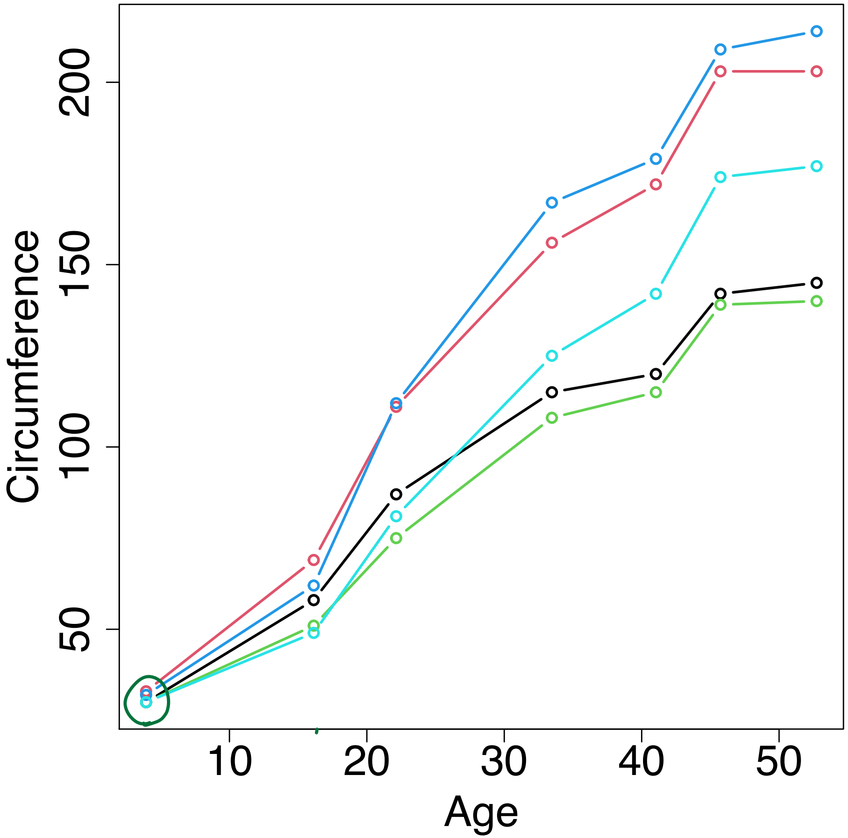
$i=2$ (p_1, p_2) (μ_1, μ_2) (σ_1^2, σ_2^2)

$i=3$, (p_1, p_2, p_3) , (μ_1, μ_2, μ_3) , $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$

• **Example 7.1.3 (Model Selection):** For 5 orange trees, the growth of tree i is measured through the circumferences y_{it} at different times T_t, resulting in the data of Table 7.1.1.

time	tree number 1	2	3	4	5
118	30	33	30	32	30
484	58	69	51	62	49
664	87	111	75	112	81
1004	115	156	108	167	125
1231	120	172	115	179	142
1372	142	203	139	209	174
1582	145	203	140	214	177

- **Example 7.1.3 (Model Selection):**



• **Example 7.1.3** (contd): The models under scrutiny are $(i = 1, \ldots, 5, t = 1, \ldots, 7)$

$$\begin{aligned}
\mathcal{M}_1 : y_{it} &\sim \mathrm{N}(\beta_{10} + b_{1i}, \sigma_1^2), \quad \times \\
\mathcal{M}_2 : y_{it} &\sim \mathrm{N}(\beta_{20} + \beta_{21} T_t + b_{2i}, \sigma_2^2), \\
\mathcal{M}_3 : y_{it} &\sim \mathrm{N}\left(\frac{\beta_{30}}{1 + \beta_{31}\exp(\beta_{32} T_t)}, \sigma_3^2\right), \quad t=0 \quad \frac{\beta_{30}}{1+\beta_{31}\cdot 1} \\
\mathcal{M}_4 : y_{it} &\sim \mathrm{N}\left(\frac{\beta_{40} + b_{4i}}{1 + \beta_{41}\exp(\beta_{42} T_t)}, \sigma_4^2\right), \quad t=0 \ , \ \frac{\beta_{40}+b_{4i}}{1+\beta_{41}}
\end{aligned}$$

where the b_{ji}'s are random effects, distributed as $\mathrm{N}(0, \tau^2)$.

† Prior modeling for model choice: Testing problem

$\sum_{\ell=1}^{k} p_\ell N(\theta_\ell, \sigma^2)$

$\lambda_j \in \{1, \ldots, k\}$

- Recall

$$\mathcal{M}_i : x \sim f_i(x \mid \theta_i), \quad \theta_i \in \Theta_i, i \in I = \{1, \ldots, p\}.$$

$Pr(\lambda_j = \ell) = p_\ell$

$Pr(M = i \mid x)$

- Assign probability p_i to the models $\mathcal{M}_i$, $i \in I$.
- Given $\mathcal{M}_i$, we define priors $\pi_i(\theta_i)$, $\theta_i \in \Theta_i$.
- Compute the posterior probability of $\mathcal{M}_i$,

$$= p(\mathcal{M}_i \mid x) = \frac{p_i m_i(x)}{\sum_j p_j m_j(x)} = \frac{p_i \int_{\Theta_i} f_i(x \mid \theta_i)\pi_i(\theta_i) d\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x \mid \theta_j)\pi_j(\theta_j) d\theta_j}.$$

$\propto p_i m_i(x)$

- Determine the model with the largest $p(\mathcal{M}_i \mid x)$.

$\propto p_i \int f_i(x \mid \theta_i) \pi_i(\theta_i) d\theta_i$

† Some difficulties: Testing problem

- Require the construction of (π_i, p_i) for each $i \in I$.
- Cannot use improper priors for π_i.

† Bayes factors (CR 7.2.2)

- Recall

$$\mathcal{M}_i : x \sim f_i(x \mid \theta_i), \quad \theta_i \in \Theta_i, i \in I = \{1, \dots, p\}.$$

- Bayes factors

$$\begin{aligned} B_{12} &= \frac{P(\mathcal{M}_1 \mid x)}{P(\mathcal{M}_2 \mid x)} \Big/ \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)} \\ &= \frac{m_1(x)}{m_2(x)} = \frac{\int_{\Theta_1} f_1(x \mid \theta_1)\pi_1(\theta_1)d\theta_1}{\int_{\Theta_2} f_2(x \mid \theta_2)\pi_2(\theta_2)d\theta_2}. \end{aligned}$$

- The model ordering is transitive; $B_{ij} = B_{ik}B_{kj}$ for $(\mathcal{M}_i, \mathcal{M}_j)$.

- Improper priors cannot be used.

† Some difficulties: Testing problem

- If some models are embedded into others, $\mathcal{M}_{i_0} \subset \mathcal{M}_{i_1}$, then there should be some coherence in the choice of π_{i_0} and π_{i_1}.

  ⋆⋆ **Example 7.1.3** (contd): Compare $\mathcal{M}_1$ and $\mathcal{M}_2$,

$$
\begin{aligned}
\mathcal{M}_1 : y_{it} &\sim \mathrm{N}(\beta_{10} + b_{1i}, \sigma_1^2), \\
\mathcal{M}_2 : y_{it} &\sim \mathrm{N}(\beta_{20} + \underbrace{\beta_{21}}_{=0} T_t + b_{2i}, \sigma_2^2). \longrightarrow \mathcal{M}_1
\end{aligned}
$$

- A larger model has more parameters to estimate with the same data ⇒ the model choice criterion must include parts that weights the fit as well as parts that incorporate the estimation error.

† Bayesian Deviance (CR 7.2.4)

$$\text{Deviance } D(\theta) = -2\log(f(x \mid \theta)).$$

$+\frac{\sum (x_i-\theta)^2}{2\sigma^2}$

- An important role in statistical model comparison
- Proportional to MSE, $1/n\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$ if the model is normal with constant variance.
- It favors higher dimensional models. $\Rightarrow$ Introduce a penalized deviance.
- For more, also see BDA §6.

$\dagger$ Deviance Information Criterion (DIC)

DIC

M_1 xx

M_2 xx $\longrightarrow$ smallest DIC

M_3 xx

$$
\begin{aligned}
\text{DIC} &= \boxed{\mathrm{E}[D(\theta) \mid x]} + p_D \\
&= \mathrm{E}[D(\theta) \mid x] + \{\mathrm{E}[D(\theta) \mid x] - D(\mathrm{E}[\theta \mid x])\} \\
&= 2\mathrm{E}[D(\theta) \mid x] - D(\underbrace{\mathrm{E}[\theta \mid x]}_{\hat{\theta}}).
\end{aligned}
$$

$\frac{1}{B} \sum_{b=1}^{B} D(\theta^{(b)}) - D(\hat{\theta})$

⋆⋆ $\mathrm{E}[D(\theta) \mid x]$: a measure of fit.

⋆⋆ p_D: a measure of model complexity (also called the effective number of parameters)

- Suggested as a criterion of model fit when the goal is to pick a model with best out-of-sample predictive power.
- Bayesian alternative to AIC and BIC.
- Allow for improper priors
- The smaller the value of DIC, the better the model

† Deviance Information Criterion (DIC) - contd

- DIC $= 2\mathrm{E}[D(\theta) \mid x] - D(\mathrm{E}[\theta \mid x])$, where $D(\theta) = -2\log(f(x \mid \theta))$.

- Given MCMC sample of $\theta^{(\ell)}$, we estimate DIC

$$
\begin{aligned}
\text{DIC} &\approx 2\hat{D}(\theta) - D(\hat{\theta}) \\
&= \frac{2}{m}\sum_{\ell=1}^{m} D(\theta^{(\ell)}) - D(\hat{\theta}),
\end{aligned}
$$

where $\hat{\theta}$ is a point estimate for θ such as the mean of the posterior simulations.

- **Example** ((PH Chapter 9) Oxygen uptake: Twelve healthy men who did not exercise regularly were recruited to take part in a study of the effects of two different exercise regimen on oxygen uptake.

$x_1 \in \{0, 1\}$

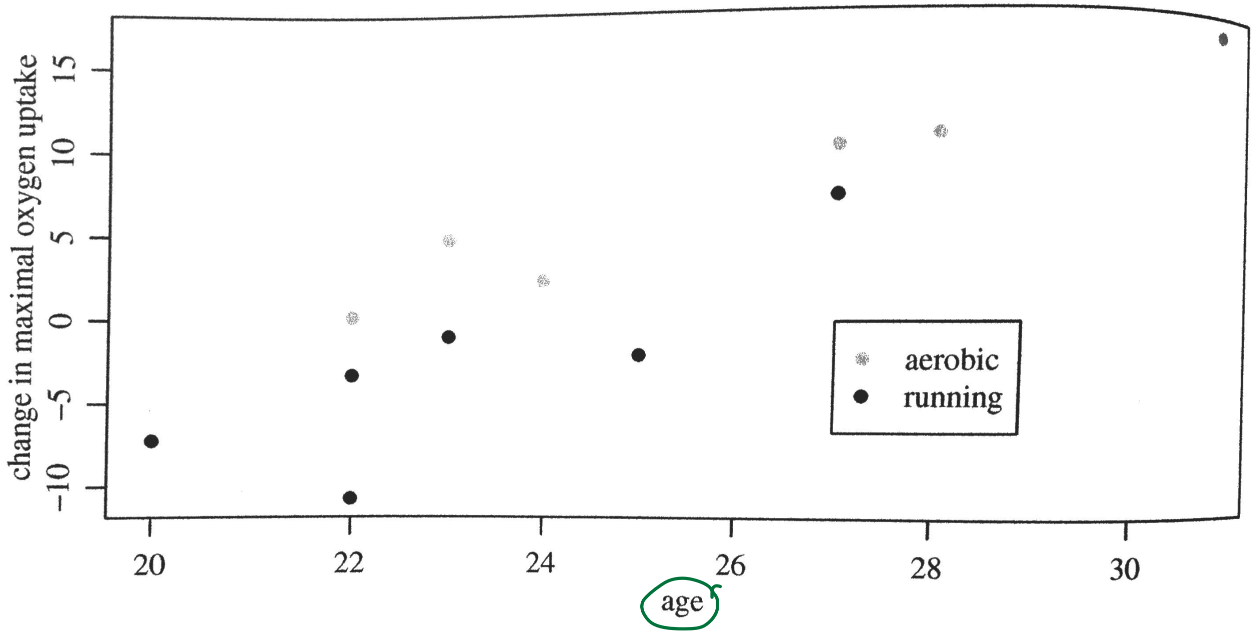
⋆⋆ Six are randomly assigned to a 12-week flat-terrain running program, and the remaining six to a 12-week step aerobics program.

⋆⋆ The maximum oxygen uptake of each subject was measured

x_2 ⋆⋆ Age is expected to affect the change in maximal uptake.

⋆⋆ Goal: want to understand how a subject's change in maximal oxygen uptake may depend on the programs.

- **Example** Oxygen uptake (contd)



- **Example** Oxygen uptake (contd)

Consider the following covariates

⋆⋆ $x_{i,1} = 0$ if subject i is on the running program, 1 if on aerobic.

⋆⋆ $x_{i,2} =$ age of subject i

⋆⋆ $x_{i,3} = x_{i,1} \times x_{i,2}$: interaction effects

- **Example** Oxygen uptake (contd)

Consider four regression model;

⋆⋆ Model 1:

$$Y_i = \beta_0 + \underline{\beta_1} \underline{x_{i,1}} + \epsilon_i, \qquad 2 \qquad \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$

⋆⋆ Model 2:

$$Y_i = \beta_0 + \beta_2 \underline{x_{i,2}} + \epsilon_i, \qquad 2 \qquad \begin{bmatrix} \beta_0 \\ \beta_2 \end{bmatrix}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_2)$ and $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$

⋆⋆ Model 3:

$$Y_i = \beta_0 + \beta_1 \underline{x_{i,1}} + \beta_2 \underline{x_{i,2}} + \epsilon_i, \qquad 3$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ and $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$ $\qquad \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$

- **Example** Oxygen uptake (contd)

Consider four regression model;

⋆⋆ Model 4:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$Y_i = \beta_0 + \beta_1 \underline{x_{i,1}} + \beta_2 \underline{x_{i,2}} + \beta_3 \underline{x_{i,3}} + \epsilon_i, \quad 4$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ and $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$

⋆⋆ Under each model, we assume

$$\pi(\underline{\boldsymbol{\beta}}, \underline{\sigma^2}) = \mathrm{N}_p(\underline{\boldsymbol{\beta}_0}, \underline{\Sigma_0})\mathrm{IG}(\underline{\nu}/2, \underline{s_0^2}/2),$$

where p denotes the number of unknown covariates. Let $\underline{\boldsymbol{\beta}_0}$, Σ_0, ν and s_0^2 fixed (HW#3-Q10(b)).

- **Example** Oxygen uptake (contd)

★★ Posterior mean estimates of the parameters;

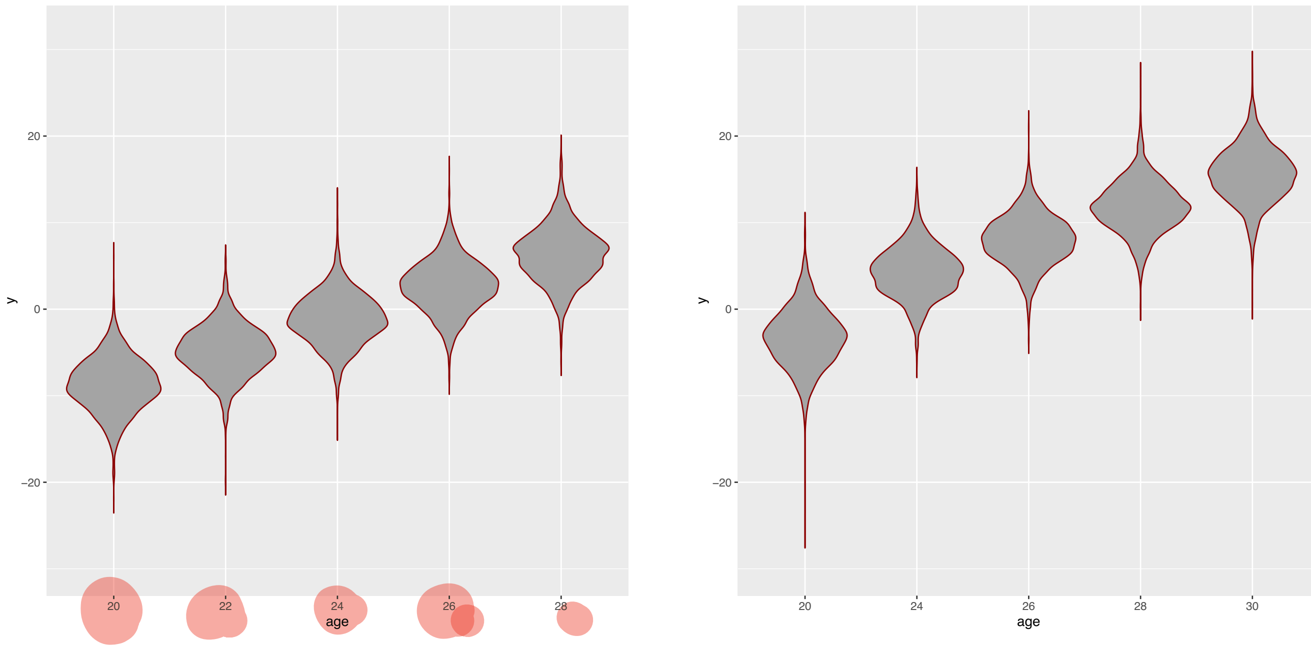
Model	β_0	β_1	β_2	β_3	σ^2	BIC
M1	-2.78	10.34			35.24	233.42
M2	-52.76		2.25		13.04	197.14
M3	-46.22	5.43	1.88		7.34	174.06
M4	-50.56	12.52	2.06	-0.289	7.86	175.79

★★ Under M3, the 95% CIs are (-59.39, -32.36), (1.95, 8.97), and (1.29, 2.45) for β_0, β_1 and β_3, respectively, and (3.135, 16.75) for σ^2

- **Example** Oxygen uptake (contd)

  ★★ Posterior predictive distributions under M3

$x_1 = 0$ $x_1 = 1$



- **Example** Bayesian model comparison: Oxygen uptake (contd)
  - ⋆⋆ We suspect many of the regression coefficients are potentially equal to zero.
  - ⋆⋆ Write the regression coefficient for variable $j = 1, 2, 3$ as $\beta_j = z_j b_j$, $z_j \in \{0, 1\}$ and $b_j \in \mathbb{R}$.

$$\Rightarrow Y_i = z_0\beta_0 + z_1\beta_1 x_{i,1} + z_2\beta_2 x_{i,2} + z_3\beta_3 x_{i,3} + \epsilon_i.$$

e.g. For $\boldsymbol{z} = (1, 0, 1, 0)$, the model is a linear regression model for y as a function of age,

$$Y_i = \beta_0 + \beta_2 \times x_{i,2} + \epsilon_i.$$

i.e., each $\boldsymbol{z}$ corresponds to a different model.

- **Example** Bayesian model comparison: Oxygen uptake (contd)
  - ⋆⋆ We place a prior distribution over $\boldsymbol{z}$, $p(\boldsymbol{z})$ and define prior distributions of non-zero β's under each $\boldsymbol{z}$.
  - ⋆⋆ Given $p(\boldsymbol{z})$, we obtain a posterior distribution over $\boldsymbol{z}$.

$$\underline{p(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{X})} = \frac{p(\boldsymbol{z})p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{z})}{\sum_{\tilde{\boldsymbol{z}}} p(\tilde{\boldsymbol{z}})p(\boldsymbol{y} \mid \boldsymbol{X}, \tilde{\boldsymbol{z}})}$$

where

$$p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{z}) = \int\int p(\boldsymbol{y} \mid \boldsymbol{z}, b, \boldsymbol{X})p(b \mid \boldsymbol{X}, \boldsymbol{z}, \sigma^2)p(\sigma^2)dbd\sigma^2.$$

p_z : # of β_js having $z_j = 1$

$\underline{\beta_z}$: vector of β_j w/ $z_j = 1$

X_z : matrix of x for j w/ $z_j = 1$

z_i

- **Example** Oxygen uptake (contd)

★★ Consider the g-prior given $\boldsymbol{z}$ and $p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{z})$ can be analytically obtained.

DE $\Rightarrow$ LASSO

$$\beta_z \mid \boldsymbol{X}_z, \sigma^2 \sim \mathrm{N}_{p_z}(\mathbf{0}, g\sigma^2(\boldsymbol{X}_z'\boldsymbol{X}_z)^{-1}).$$

and assume $\sigma^2 \sim$ IG and g fixed at n.

$\boldsymbol{z}$	model	$\log p(\boldsymbol{y}\mid\mathbf{X}, \boldsymbol{z})$	$p(\boldsymbol{z}\mid\boldsymbol{y}, \mathbf{X})$
(1,0,0,0)	β_1	-44.33	0.00
(1,1,0,0)	$\beta_1 + \beta_2 \times \text{group}_i$	-42.35	0.00
(1,0,1,0)	$\beta_1 + \beta_3 \times \text{age}_i$	-37.66	0.18
(1,1,1,0)	$\beta_1 + \beta_2 \times \text{group}_i + \beta_3 \times \text{age}_i$	-36.42	0.63
(1,1,1,1)	$\beta_1 + \beta_2 \times \text{group}_i + \beta_3 \times \text{age}_i + \beta_4 \times \text{group}_i \times \text{age}_i$	-37.60	0.19

Table 9.1. Marginal probabilities of the data under five different models.

$\underbrace{\boldsymbol{z} = (z_1, z_2, z_3, z_4)}$ $z_j \sim \mathrm{Ber}(p_j)$

$P(\boldsymbol{z}) = \prod p_j^{z_j} (1-p_j)^{1-z_j}$