# Part 6:

# Multivariate Normal and Linear Models

# Multiple measurements

- Up until now all of our statistical models have been univariate models
  - models for a single measurement on each member of a sample of individuals, or each run of a repeated experiment
- However, datasets are frequently multivariate, having multiple measurements for each individual or experiment

# Multivariate model

- The most useful (commonly used) model for multivariate data is the multivariate normal model
- For a collection of variables, it allows us to jointly estimate population
  - means
  - variances
  - and correlations

# Example: reading comprehension

A sample of 22 children are given reading comprehension tests before and after receiving a particular instructional method

Each student $i$ will have two scores, $Y_{i,1}$ and $Y_{i,2}$ denoting the pre- and post-instructional scores respectively

We denote each student's pair of scores as a vector $\mathbf{Y}_i$, so that

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} = \begin{pmatrix} \text{score on first test} \\ \text{score on second test} \end{pmatrix}$$

# Example: quantities of interest

The things we might be interested in include the population mean $\mu$, particularly $\mu_2 - \mu_1$

$$\mathbb{E}\{\mathbf{Y}\} = \begin{pmatrix} \mathbb{E}\{Y_{i,1}\} \\ \mathbb{E}\{Y_{i,2}\} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

and the population covariance matrix $\Sigma$

$$\Sigma = \mathrm{Cov}[\mathbf{Y}]$$

$$= \begin{pmatrix} \mathbb{E}\{Y_1^2\} - \mathbb{E}\{Y_1\}^2 & \mathbb{E}\{Y_1 Y_2\} - \mathbb{E}\{Y_1\}\mathbb{E}\{Y_2\} \\ \mathbb{E}\{Y_1 Y_2\} - \mathbb{E}\{Y_1\}\mathbb{E}\{Y_2\} & \mathbb{E}\{Y_2^2\} - \mathbb{E}\{Y_2\}^2 \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}$$

[$\rho$ measures the consistency of the intervention]

where $\sigma_{1,2} = \rho\sigma_1\sigma_2$

for $\rho \in [0, 1]$

# Multivariate normal

One model for describing first- and second-order moments of multivariate data is the multivariate normal model

We say that a $p$-dimensional data vector $\mathbf{Y}$ has a multivariate normal (MVN) distribution if its sampling density is

$$p(\mathbf{y}|\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mu)^\top \Sigma^{-1}(\mathbf{y}-\mu)\right\}$$

where

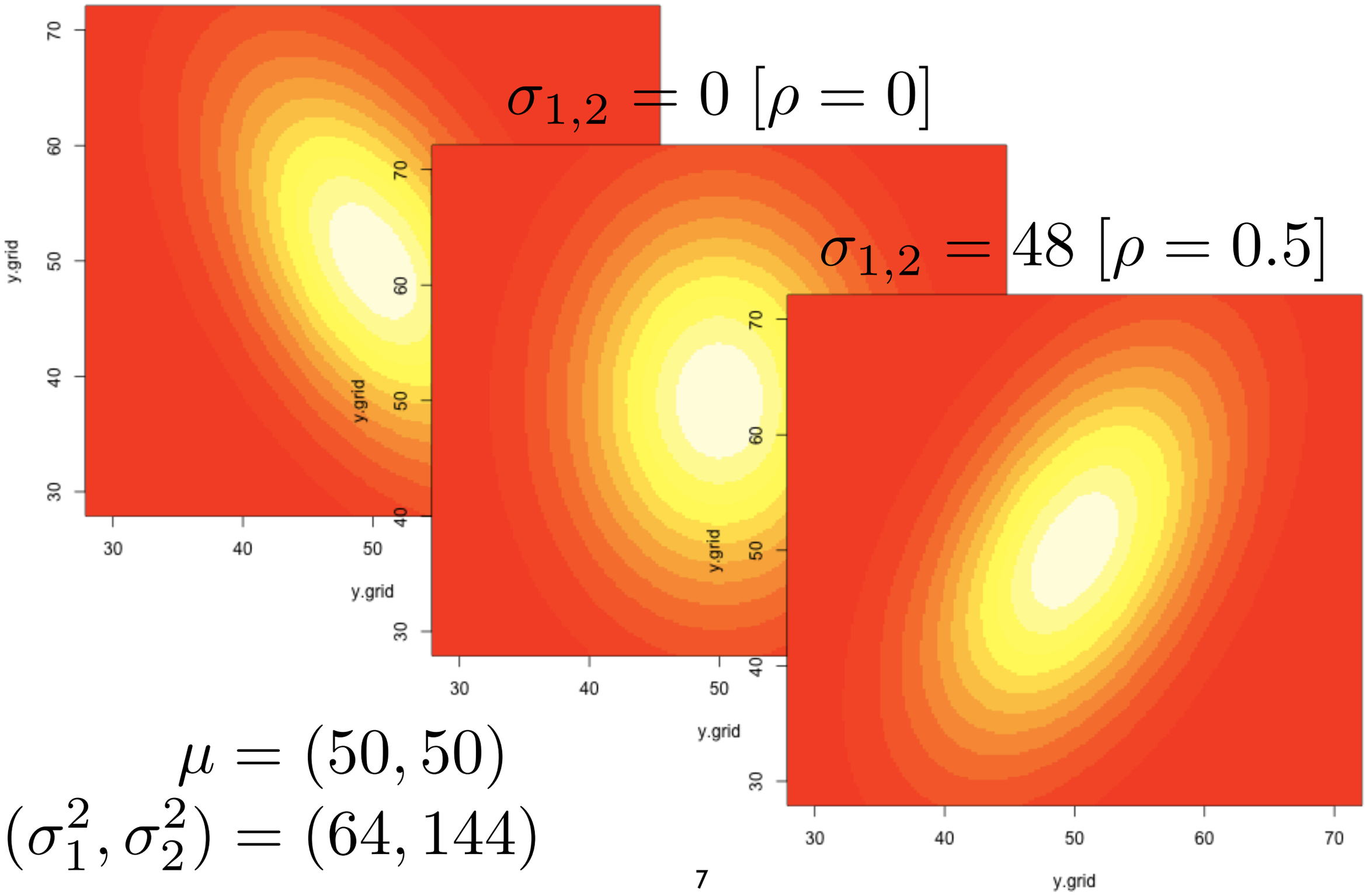
$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1,p} \\ \vdots & \ddots & \vdots \\ \sigma_{1,p} & \cdots & \sigma_p^2 \end{pmatrix}$$

# Example: bivariate normals

$$\sigma_{1,2} = -48 \; [\rho = -0.5]$$

$$\sigma_{1,2} = 0 \; [\rho = 0]$$

$$\sigma_{1,2} = 48 \; [\rho = 0.5]$$




$$\mu = (50, 50)$$

$$(\sigma_1^2, \sigma_2^2) = (64, 144)$$

# Notation & marginals

We shall write

$$\mathbf{Y} \sim \mathcal{N}_p(\mu, \Sigma)$$

when $\mathbf{Y}$ has a $p$-dimensional MVN distribution

An interesting feature of the MVN distribution is that the marginal distribution of each variable is a univariate normal:

$$Y_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

# Semiconjugate prior

Recall that if $Y_1, \ldots, Y_n$ are IID (univariate) normal, then a convenient conjugate prior distribution for the population mean is also (univariate) normal

Similarly, a convenient prior distribution for the multivariate mean $\mu$ is a MVN distribution, which we will parameterize as

$$p(\mu) = \mathcal{N}_p(\mu; \mu_0, \Lambda_0)$$

where $\mu_0$ and $\Lambda_0$ are the prior mean and variance of $\mu$, respectively

# The essence of the prior

We have that if $\mu \sim \mathcal{N}_p(\mu_0, \Lambda_0)$, then

$$p(\mu) \propto \exp\left\{-\frac{1}{2}\mu^\top A_0 \mu + \mu^\top b_0\right\}$$

where $A_0 = \Lambda_0^{-1}$ and $b_0 = \Lambda_0^{-1}\mu_0$

Conversely, this result says that if a random vector $\mu$ has a density in $\mathbb{R}^p$ that is proportional to

$$\exp\left\{-\frac{1}{2}\mu^\top A \mu + \mu^\top b\right\}$$

for some matrix $A$ and vector $b$, then $\mu$ must have a MVN distribution with covariance $A^{-1}$ and mean $A^{-1}b$

# The essence of likelihood

If the sampling model is

$$\{\mathbf{Y}_1, \ldots, \mathbf{Y}_n | \mu, \Sigma\} \overset{\text{iid}}{\sim} \mathcal{N}_p(\mu, \Sigma)$$

then similar calculations show that the joint sampling density of the observed vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ is

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_n | \mu, \Sigma) \propto \exp\left\{ -\frac{1}{2}\mu^\top A_1 \mu + \mu^\top b_1 \right\}$$

where $A_1 = n\Sigma^{-1}, b_1 = n\Sigma^{-1}\bar{\mathbf{y}}$ and

$$\bar{\mathbf{y}} = (\tfrac{1}{n}\sum y_{i,1}, \ldots, \tfrac{1}{n}\sum y_{i,n})$$

# The essence of posterior

Combining the prior and likelihood gives the posterior

$$p(\mu|\mathbf{y}_1,\ldots,\mathbf{y}_n,\Sigma) \propto \exp\left\{\frac{1}{2}\mu^\top A_1\mu + \mu^\top b_1\right\}$$

$$\times \exp\left\{\frac{1}{2}\mu^\top A_0\mu + \mu^\top b_0\right\}$$

$$= \exp\left\{\frac{1}{2}\mu^\top A_n\mu + \mu^\top b_n\right\}$$

where

$$A_n = A_0 + A_1 = \Lambda_0^{-1} + n\Sigma^{-1}$$

$$b_n = b_0 + b_1 = \Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{\mathbf{y}}$$

# The posterior

This implies that the posterior conditional distribution of $\mu$ must therefore be MVN with covariance $A_n^{-1}$ and mean $A_n^{-1} b_n$, so

$$\mathrm{Cov}[\mu|\mathbf{y}_1, \ldots, \mathbf{y}_n, \Sigma] = \Lambda_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}$$

$$\mathbb{E}\{\mu|\mathbf{y}_1, \ldots, \mathbf{y}_n, \Sigma\} = \mu_n = \Lambda_n(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{\mathbf{y}})$$

giving $\{\mu|\mathbf{y}_1, \ldots, \mathbf{y}_n, \Sigma\} \sim \mathcal{N}_p(\mu_n, \Lambda_n)$

Just like in the univariate case:

- the posterior precision (inverse covariance) is the sum of the prior precision and data precision
- the posterior expectation is the average of the prior expectation and the sample mean

# Prior covariance elements

Just as the variance $\sigma^2$ must be positive, the covariance matrix $\Sigma$ must be positive definite, i.e.,

$$x^\top \Sigma x > 0, \quad \text{for all vectors} \;\; x$$

Positive definiteness guarantees that $\sigma_j^2 > 0$ for all $j$ and that all correlations are between $-1$ and $1$

Another requirement is that the covariance matrix must be symmetric, i.e., $\sigma_{j,k} = \sigma_{k,j}$

Any valid prior distribution for $\Sigma$ must put all of its probability mass on this complicated set of symmetric, positive definite matrices

# Wishart distribution

A convenient family of distributions with just these properties is the Wishart

- the multivariate analogue of the gamma family

Recall that, in the univariate normal model, the gamma distribution is conjugate for the *precision* $1/\sigma^2$

- the conjugate prior for $\sigma^2$ is IG

Similarly, it turns out that the Wishart distribution is a semi-conjugate prior for the *precision matrix* $\Sigma^{-1}$

- and so the conjugate prior for $\Sigma$ is inverse-Wishart

# Inverse-Wishart

To obtain $\Sigma \sim \mathrm{IW}(\nu_0, S_0^{-1})$ where $S_0$ is a $p \times p$ positive-definite matrix and $\nu_0$ is a positive integer

1. Sample $\mathbf{z}_1, \ldots, \mathbf{z}_{\nu_0} \overset{\text{iid}}{\sim} \mathcal{N}_p(\mathbf{0}, S_0^{-1})$

2. Calculate $Z^\top Z = \sum_{i=1}^{\nu_0} \mathbf{z}_i \mathbf{z}_i^\top$

3. Set $\Sigma = (Z^\top Z)^{-1}$

Accordingly, the *precision matrix* $\Sigma^{-1}$ has a $W(\nu_0, S_0^{-1})$ distribution

# Expected covariance

The expected covariances under an inverse-Wishart are

$$\mathbb{E}\{\Sigma^{-1}\} = \nu_0 S_0^{-1}$$

$$\mathbb{E}\{\Sigma\} = \frac{1}{\nu - p - 1} S_0$$

As a prior for a MVN covariance, if we are confident that the true $\Sigma$ is near $\Sigma_0$, then we might choose $\nu_0$ large and set $S_0 = (\nu_0 - p - 1)\Sigma_0$ so that the distribution is tightly centered around $\Sigma_0$

If not, we may choose $\nu_0 = p + 2$ and $S_0 = \Sigma_0$, so that the distribution is loosely centered around $\Sigma_0$

# IW (prior) density

The density for $\mathrm{IW}(\nu_0, S_0^{-1})$ is given by

$$p(\Sigma) \propto |\Sigma|^{-(\nu_0+p+1)/2} \times \exp\left\{-\frac{1}{2}\mathrm{tr}(S_0\Sigma^{-1})\right\}$$

where $\mathrm{tr}(\cdot)$ is the trace, or sum of the diagonal elements, of a matrix

An interesting result from linear algebra is that

$$\sum_{k=1}^{K} b_k^\top A b_k = \mathrm{tr}(BAB^\top) = \mathrm{tr}(B^\top BA)$$

where $B$ is a matrix whose $k^{\mathrm{th}}$ row is $b_k^\top$

# Convenient likelihood

To combine the IW prior distribution with the sampling distribution for $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ we take advantage of the above result for traces:

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_n | \mu, \Sigma)$$

$$\propto |\Sigma|^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\mathbf{y}_i - \mu^\top) \Sigma^{-1} (\mathbf{y}_i - \mu) \right\}$$

$$= |\Sigma|^{-n/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}(S_\mu \Sigma^{-1}) \right\}$$

where $S_\mu = \sum_{i=1}^{n} (\mathbf{y}_i - \mu)(\mathbf{y}_i - \mu)^\top$ is the residual sum of squares matrix for the vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ if the population mean is presumed to be $\mu$

# Posterior

We are now in a convenient position to combine the prior with the likelihood as follows

$$
\begin{aligned}
& p(\Sigma|\mathbf{y}_1, \ldots, \mathbf{y}_n, \mu) \\
& \quad \propto p(\mathbf{y}_1, \ldots, \mathbf{y}_n|\mu, \Sigma) \times p(\Sigma) \\
& = \left( |\Sigma|^{-n/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}(S_\mu \Sigma^{-1}) \right\} \right) \\
& \times \left( |\Sigma|^{-(\nu_0+p+1)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}(S_0 \Sigma^{-1}) \right\} \right) \\
& = |\Sigma|^{-(\nu_0+n+p+1)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}([S_0 + S_\mu]\Sigma^{-1}) \right\}
\end{aligned}
$$

# Posterior

Thus we have

$$\{\Sigma|\mathbf{y}_1,\ldots,\mathbf{y}_n,\mu\} \sim \mathrm{IW}(\nu_0+n,[S_0+S_\mu]^{-1})$$

Even though it was rough going to obtain, hopefully the result is intuitive

- the posterior sample size $\nu_n = \nu_0 + n$ is the sum of the prior sample size and the data sample size
- Similarly, the posterior residual sum of squares $S_n = S_0 + S_\mu$ is a sum of the prior sum of squares and the data sum of squares

# Gibbs sampler

We now have the set of full posterior conditionals

$$\{\mu|\mathbf{y}_1,\ldots,\mathbf{y}_n,\Sigma\} \sim \mathcal{N}_p(\mu_n,\Lambda_n)$$

$$\{\Sigma|\mathbf{y}_1,\ldots,\mathbf{y}_n,\mu\} \sim \mathrm{IW}(\nu_n,S_n^{-1})$$

which we may use to construct a GS algorithm to obtain a MCMC approximation to the joint posterior

$$p(\mu,\Sigma|\mathbf{y}_1,\ldots,\mathbf{y}_n)$$

# Gibbs sampler

Given a starting value $\Sigma^{(0)}$, the GS algorithm generates $\{\mu^{(s+1)}, \Sigma^{(s+1)}\}$ from $\{\mu^{(s)}, \Sigma^{(s)}\}$ via the following two steps:

1. Sample $\mu^{(s+1)}$ from its full conditional distribution:
   a) compute $\mu_n$ and $\Lambda_n$ from $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and $\Sigma^{(s)}$
   b) sample $\mu^{(s+1)} \sim \mathcal{N}_p(\mu_n, \Lambda_n)$

2. Sample $\Sigma^{(s+1)}$ from its full conditional distribution:
   a) compute $S_n$ from $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and $\mu^{(s+1)}$
   b) sample $\Sigma^{(s+1)} \sim \mathrm{IW}(\nu_n, S_n^{-1})$

a) $\Rightarrow \{\mu_n, \Lambda_n\}$ depend on $\Sigma$ and $S_0$ depends on $\mu$, so these must be recalculated every iteration

# Example: reading comprehension

Lets return to the our motivating reading comprehension example

We shall model the 22 pairs of scores, before and after instruction, as IID samples from a MVN distribution

We start by thinking about the priors for $\mu$ and $\Sigma$

The exam was designed to give average scores of around 50 out of 100, so $\mu_0 = (50, 50)$ is sensible as a prior expectation

# Example: mean prior variance

Since the true mean cannot be below 0 or above 100, it is desirable to use a prior variance for $\mu$ that puts little probability outside of this range

If we take $\lambda_{0,1}^2 = \lambda_{0,2}^2 = (50/2)^2 = 625$ then $P(\mu_j \notin [0, 100]) = 0.05$

Since the two exams are measuring the same thing, whatever the true values of $\mu_1$ and $\mu_2$ are, it is probable that they are close

We can reflect this with a prior correlation of 0.5, so that $\lambda_{1,2} = 312.5$

# Example: prior covariance & data

Some of the same logic about the range of exam scores applies to choosing a prior for $\Sigma$

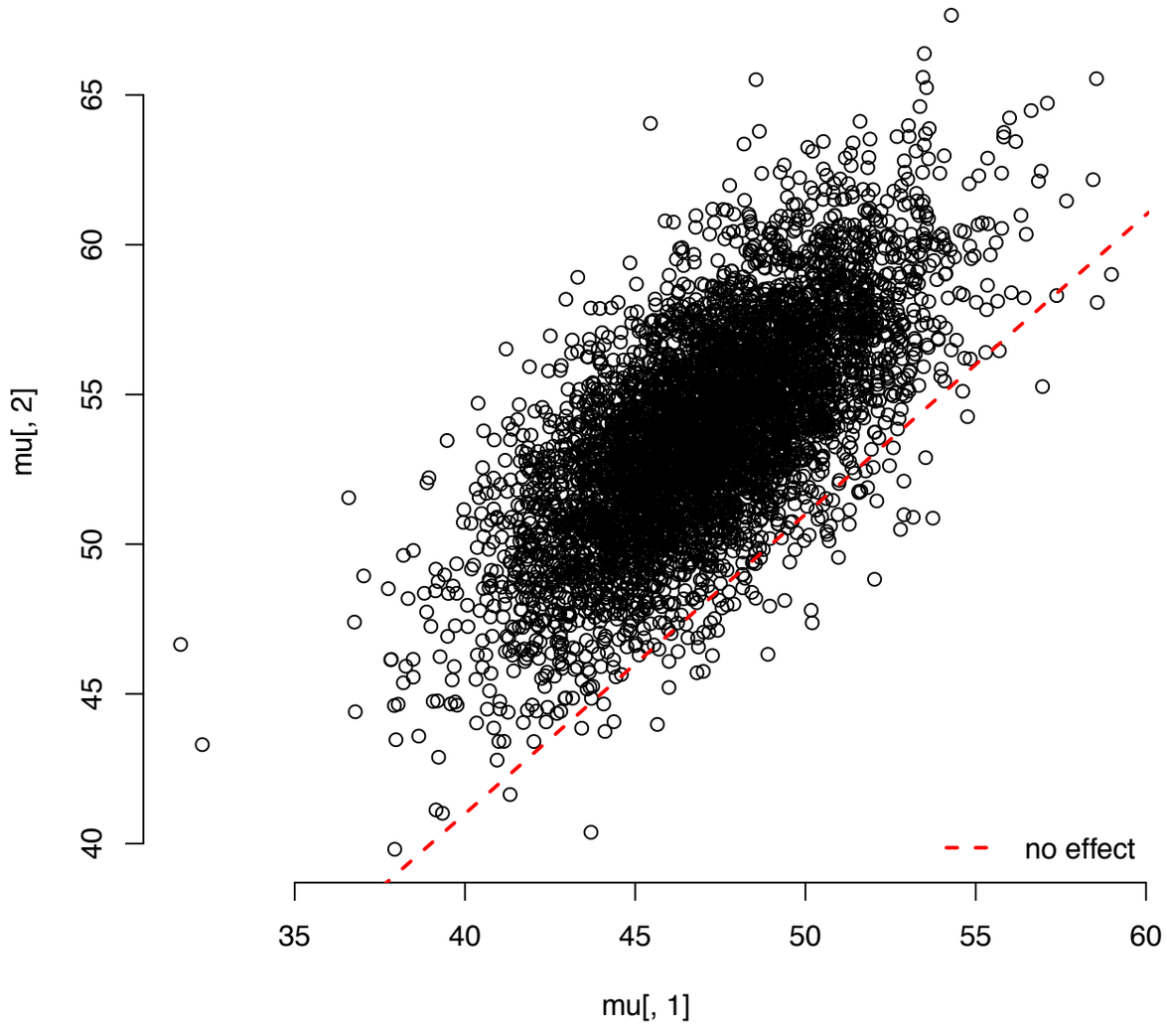
We'll take $S_0 = \Lambda_0$, but only loosely center $\Sigma$ around this value by taking $\nu_0 = p + 2 = 4$

The sufficient statistics of $\mathbf{y}_1, \ldots, \mathbf{y}_{22}$ needed for MVN inference are

$$\begin{aligned} \bar{\mathbf{y}} &= (47.18, 53.86)^\top \\ (s_1^2, s_2^2) &= (182.16, 243.65) \\ s_{1,2}/(s_1 s_2) &= 0.7 \end{aligned}$$

# Example: Gibbs sampler

Initialize the GS algorithm with $\Sigma^{(0)} = \begin{pmatrix} s_1^2 & s_{1,2} \\ s_{1,2} & s_2^2 \end{pmatrix}$




## Quantile-based CIs

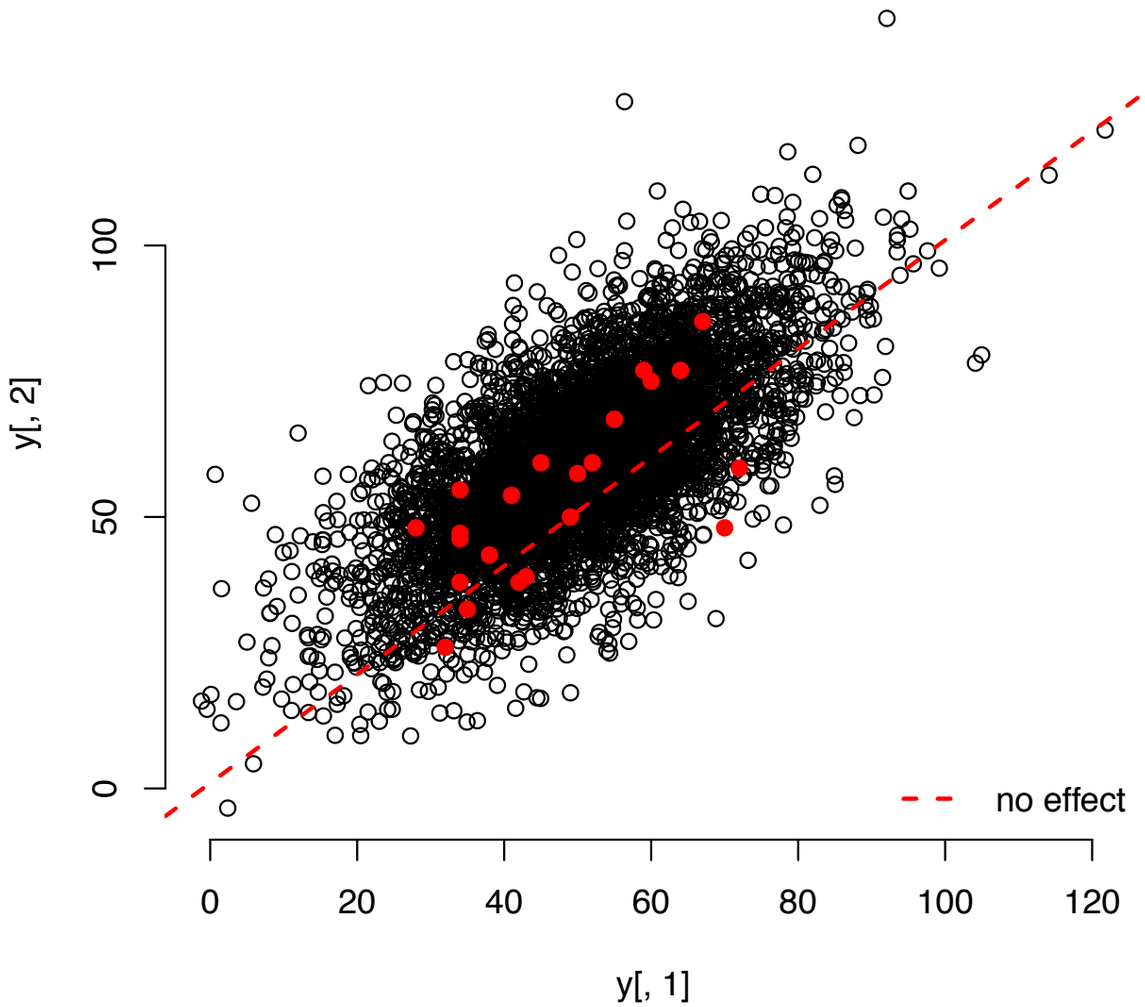
| | 2.5% | 97.5% |
|---|---|---|
| $\mu_2 - \mu_1$ | 1.48 | 11.74 |

$$P(\mu_2 > \mu_1 | \mathbf{y}_1, \ldots, \mathbf{y}_{22}) > 0.99$$

strong evidence that the instruction is working

# Example: Posterior predictive

What is the probability that a randomly selected child will score higher on the second exam than on the first?




## Quantile-based CIs

| | 2.5% | 97.5% |
|---|---|---|
| $y_2 - y_1$ | -14.01 | 34 |

$$P(y_2 > y_1 | \mathbf{y}_1, \ldots, \mathbf{y}_{22}) = 0.78$$

a less significant, and possibly worrying result!

# Regression

Regression modeling is concerned with describing how the sampling distribution of one random variable

$Y$ response variable

varies with another variable or set of variables

$x = (x_1, \ldots, x_p)$ explanatory variable(s) or "covariates"

Specifically, a regression model postulates a form for $p(y|x)$, the conditional distribution for $Y$ given $x$

Estimation of $p(y|x)$ is made using data $y_1, \ldots, y_n$ gathered under a variety of conditions $x_1, \ldots, x_n$

# Linear model

One simple but flexible approach to regression is via the linear (sampling) model (LM)

The LM treats responses $Y_i$ as independent (but not identically distributed) realizations of a process that is linear in explanatory variables $x_i^\top = (x_{i,1}, \ldots, x_{i,p})$, observed with Gaussian noise

$$Y_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma^2) \quad \text{where} \quad \mu_i = \mathbb{E}\{Y|x_i\} = x_i^\top \beta$$

where the $x_i$ are known, $\beta$ is an unknown $p$-dimensional parameter vector of regression coefficients, and $\sigma^2$ is an unknown variance parameter

# Compact notation

The LM is usually written as $Y = X\beta + \varepsilon$, where

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

and $\{\varepsilon_1, \ldots, \varepsilon_n\} \overset{\text{iid}}{\sim} N(0, \sigma^2)$

Even more compact notation is

design matrix

$$Y \sim \mathcal{N}_n(X\beta, I_n\sigma^2)$$

where $I_n$ is a $n \times n$ identity matrix

# Maximum Likelihood

As long as $X^\top X$ is invertible, which is the case when $X$ is of full rank $p$, the MLE is

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

Similarly, we may find

$$\hat{\sigma}^2 = \frac{1}{n}||Y - X\hat{\beta}||^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - x_i^\top \hat{\beta})^2$$

# Sampling the MLE

- It may be shown that $\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(X^\top X)^{-1})$

and $\hat{\sigma}^2 \sim \frac{\sigma^2}{n}\chi^2_{n-p}$

- Moreover $\hat{\beta}$ and $\hat{\sigma}^2$ are independent

These results may be used to construct confidence intervals, test hypotheses, etc., in a frequentist setup

# Example: Oxygen intake

12 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 intake

- 6 were randomly assigned to a 12-week flat-terrain running program
- the remaining 6 were assigned a 12-week step aerobics program

The maximum oxygen intake of each subject was measured (in l/m) while running on an inclined treadmill, both before and after the 12-week program

# Example: Oxygen intake

Of interest is how a subject's change in maximal oxygen intake may depend upon which program they were assigned to

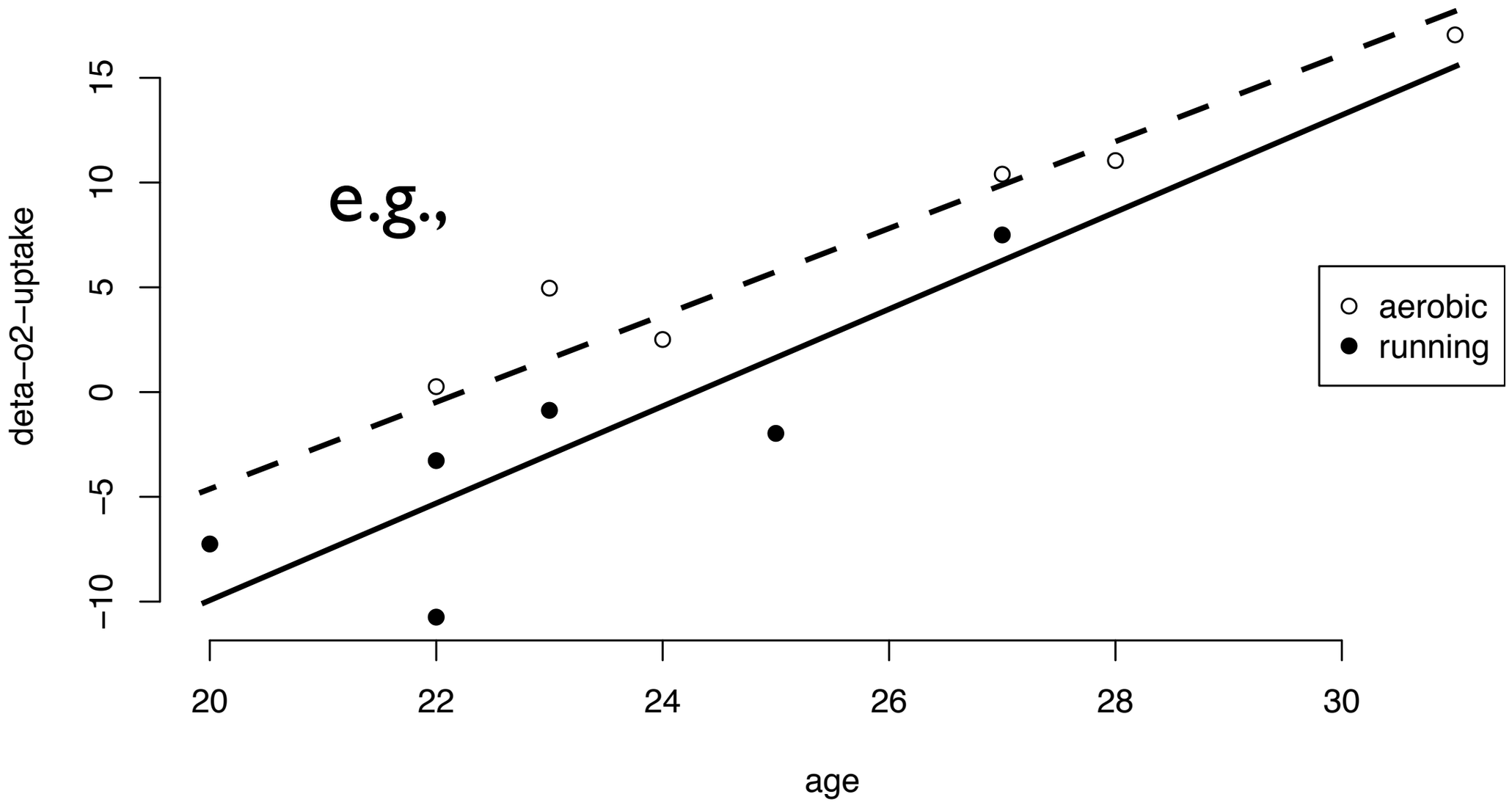
- one explanatory variable

However, other factors, such as age, are expected to affect the change in maximal intake as well

- two explanatory variables

I.e., we wish to estimate the conditional distribution of oxygen intake for a given exercise program and age

# Example: Data




It is easy to "imagine" two straight lines, one for the aerobic points, and the for the running ones

How do we estimate them, and do we need two?

# Example: Linear in "covariates"

A sensible LM may be constructed as follows

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4}$$

$x_{i,1} = 1$ for each subject $i$ (intercept term)

$x_{i,2} = 0$ if subject $i$ is running, 1 if aerobic

$x_{i,3} =$ the age of subject $i$

$x_{i,4} = x_{i,2} \times x_{i,3}$ (interaction term)

Thus, the (12) rows of $X$ are comprised of 4 columns:

$$x_i^\top = (x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4})$$

# Example: Explaining the terms in the model

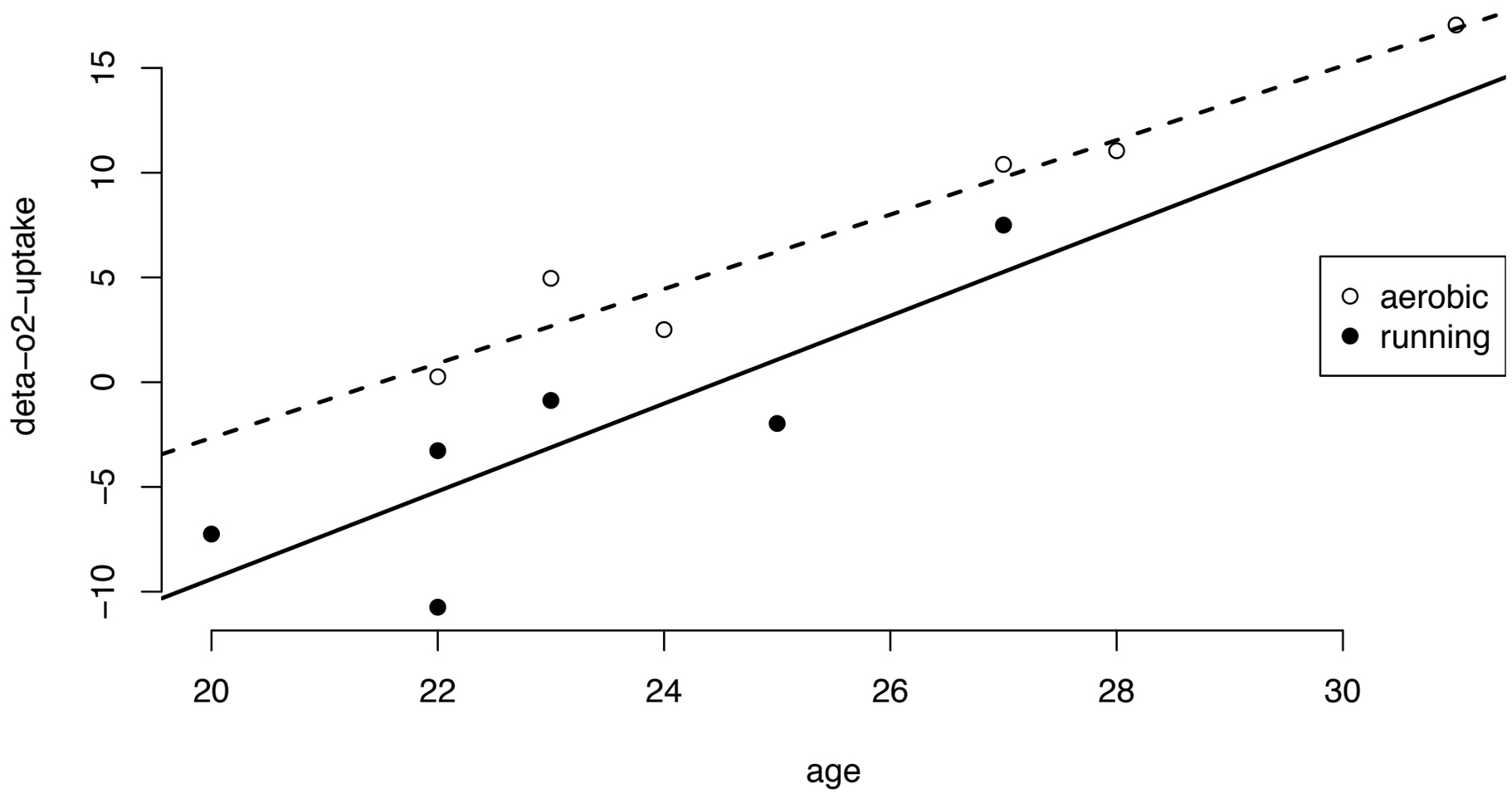
Under this model the conditional expectations of $Y$ for the two different levels of $x_{i,2}$ are

$$\mathbb{E}\{Y|x\} = \beta_1 + \beta_3 \times \text{age}, \quad \text{if running}$$
$$\mathbb{E}\{Y|x\} = (\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times \text{age}, \quad \text{if aerobic}$$

The model assumes that the relationship is linear in age for both exercise groups, with

- the difference in intercepts given by $\beta_2$
- and the difference in slopes by $\beta_4$

# Example: MLE/OLS inference




Classical tests indicate that the exercise program may not be significant

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| x1 | -51.2939 | 12.2522 | -4.187 | 0.00305 | ** |
| x2 | 13.1071 | 5.7620 | 0.832 | 0.42978 | |
| x3 | 2.0947 | 0.5264 | 3.980 | 0.00406 | ** |
| x4 | -0.3182 | 0.6498 | -0.490 | 0.63746 | |

# Bayesian LM

We demonstrate how a simple semi-conjugate prior distribution for $\beta$ and $\sigma^2$ can be used when there is information available about the parameters

In situations where prior information is unavailable or difficult to quantify, an alternative "default" class of prior distributions is given

We shall see how the MLE $(\hat{\beta}, \hat{\sigma}^2)$ crops up as factors in our posterior distributions, with similar sampling distributions as posteriors in the default prior case

# Semi-conjugate prior

The sampling density of the data, as a function of $\beta$ is

$$p(y|X,\beta,\sigma^2) = \mathcal{N}_n(y; X\beta, \sigma^2)$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^\top (y - X\beta)\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}[y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta]\right\}$$

The role that $\beta$ plays in the exponent looks very similar to that played by $y$, which is MVN

This suggests that a MVN prior for $\beta$ is conjugate

# Semi-conjugate prior

If $\beta \sim \mathcal{N}_p(\beta_0, \Sigma_0)$, then

$$p(\beta|y, X, \sigma^2) \propto$$
$$\exp\left\{\beta^\top\left(\Sigma_0^{-1}\beta_0 + \frac{X^\top y}{\sigma^2}\right) - \frac{1}{2}\beta^\top\left(\Sigma_0^{-1} + \frac{X^\top X}{\sigma^2}\right)\beta\right\}$$

which we recognize as proportional to an MVN with

$$\Sigma_n \equiv \mathrm{Var}[\beta|y, X, \sigma^2] = (\Sigma_0^{-1} + X^\top X/\sigma^2)^{-1}$$
$$\beta_n \equiv \mathbb{E}\{\beta|y, X, \sigma^2\} = \Sigma_n(\Sigma_0^{-1}\beta_0 + X^\top y/\sigma^2)$$

i.e., $\{\beta|y, X, \sigma^2\} \sim \mathcal{N}_p(\beta_n, \Sigma_n)$

# Interpretation

We can gain some understanding of the posterior (conditional) by considering some limiting cases of

$$\Sigma_n \equiv \mathrm{Var}[\beta|y, X, \sigma^2] = (\Sigma_0^{-1} + X^\top X/\sigma^2)^{-1}$$

$$\beta_n \equiv \mathbb{E}\{\beta|y, X, \sigma^2\} = \Sigma_n(\Sigma_0^{-1}\beta_0 + X^\top y/\sigma^2)$$

If the elements of the prior precision matrix $\Sigma_0^{-1}$ are small, then $\beta_n \approx \hat{\beta}$, the MLE (or OLS estimator)

On the other hand, if the measurement precision is very small ($\sigma^2$ is very large), then $\beta_n \approx \beta_0$, the prior expectation

# Conjugate prior variance

As in most normal sampling problems, the semi-conjugate prior for $\sigma^2$ is IG

If $\sigma^2 \sim \mathrm{IG}(\nu_0/2, \nu_0\sigma_0^2/2)$ then

$$
\begin{aligned}
&p(\sigma^2|yX,\beta) \\
&\quad = (\sigma^2)^{-(\nu_0+n)/2+1} \exp\left\{-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + (y-X\beta)^\top(y-X\beta)]\right\}
\end{aligned}
$$

which we recognize as an IG density, i.e.,

$$
\{\sigma^2|y,X,\beta\} \sim \mathrm{IG}\left(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2 + (y-X\beta)^\top(y-X\beta)}{2}\right)
$$

# Gibbs sampler

Using these full conditionals, we may construct a GS algorithm as follows: given current values of $\{\beta^{(s)}, \sigma^{2(s)}\}$, new values can be generated by

1. updating $\beta$

   a) compute $\Sigma_n(y, X, \sigma^{2(s)})$ and $\beta_n(y, X, \sigma^{2(s)})$
   b) sample $\beta^{(s+1)} \sim \mathcal{N}_p(\beta_n, \Sigma_n)$

2. updating $\sigma^2$

   a) compute $s^2_\beta = (y - X\beta^{(s+1)})^\top (y - X\beta^{(s+1)})$
   b) sample $\sigma^{2(s+1)} \sim \mathrm{IG}([\nu_0 + n]/2, [\nu_0\sigma_0^2 + s^2_\beta]/2)$

# Prior difficulty

The Bayesian analysis of the LM requires a specification of the prior parameters $(\beta_0, \Sigma_0)$ and $(\nu_0, \sigma_0^2)$

There are $O(p^2)$ such parameters, so this can be quite a monumental task even for modest $p$

Even when prior information exists (as in the oxygen intake example) sometimes an analysis must be done in the absence of prior information

Fortunately, there are some convenient weakly-informative priors that are easy to use

# Jeffreys' prior

The (independence) Jeffreys prior for the LM is

$$p(\beta, \sigma^2) \propto 1/\sigma^2$$

i.e., $p(\beta|\sigma) = p(\beta) \propto 1$ and $p(\sigma^2) \propto 1/\sigma^2$

We may interpret the former as $\beta \sim \mathcal{N}(\mathbf{0}, \infty_p)$, and the latter as $\sigma^2 \sim \mathrm{IG}(0, 0)$, from which we may easily derive

$$\beta|y, X, \sigma^2 \sim \mathcal{N}_p(\hat{\beta}, \sigma^2(X^\top X)^{-1})$$

$$\sigma^2|y, X, \beta \sim \mathrm{IG}(n/2, n\hat{\sigma}^2/2)$$

Check that the posterior is proper for $n > p + 1$

# Marginal posterior variance

The marginal posterior variance is available in closed form under the IG (and Jeffreys') prior

$$p(\sigma^2|y,X) \overset{\text{(cond. prob)}}{=} \frac{p(\sigma^2,\beta|y,X)}{p(\beta|\sigma^2,y,X)}$$

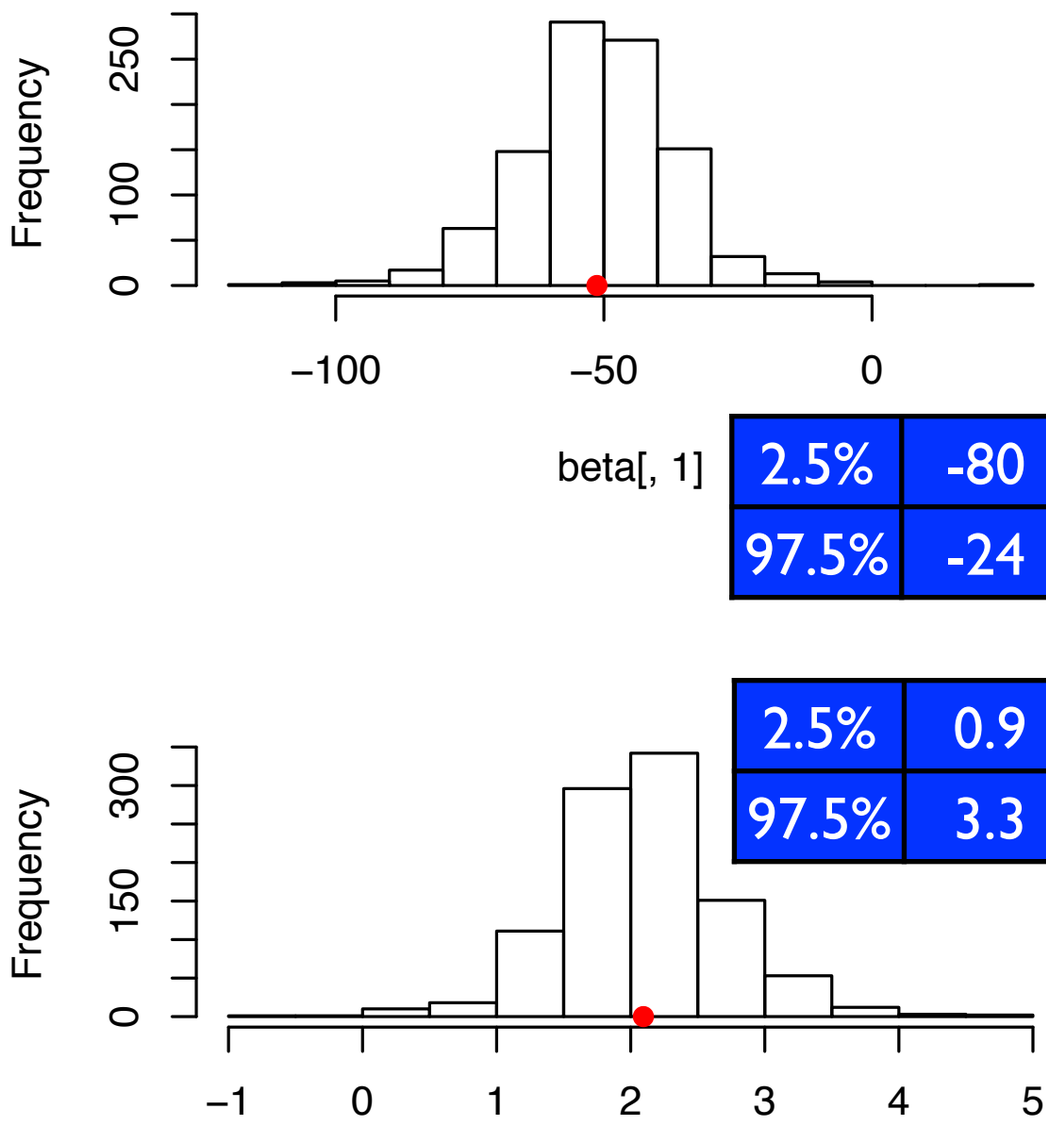
$$\vdots$$

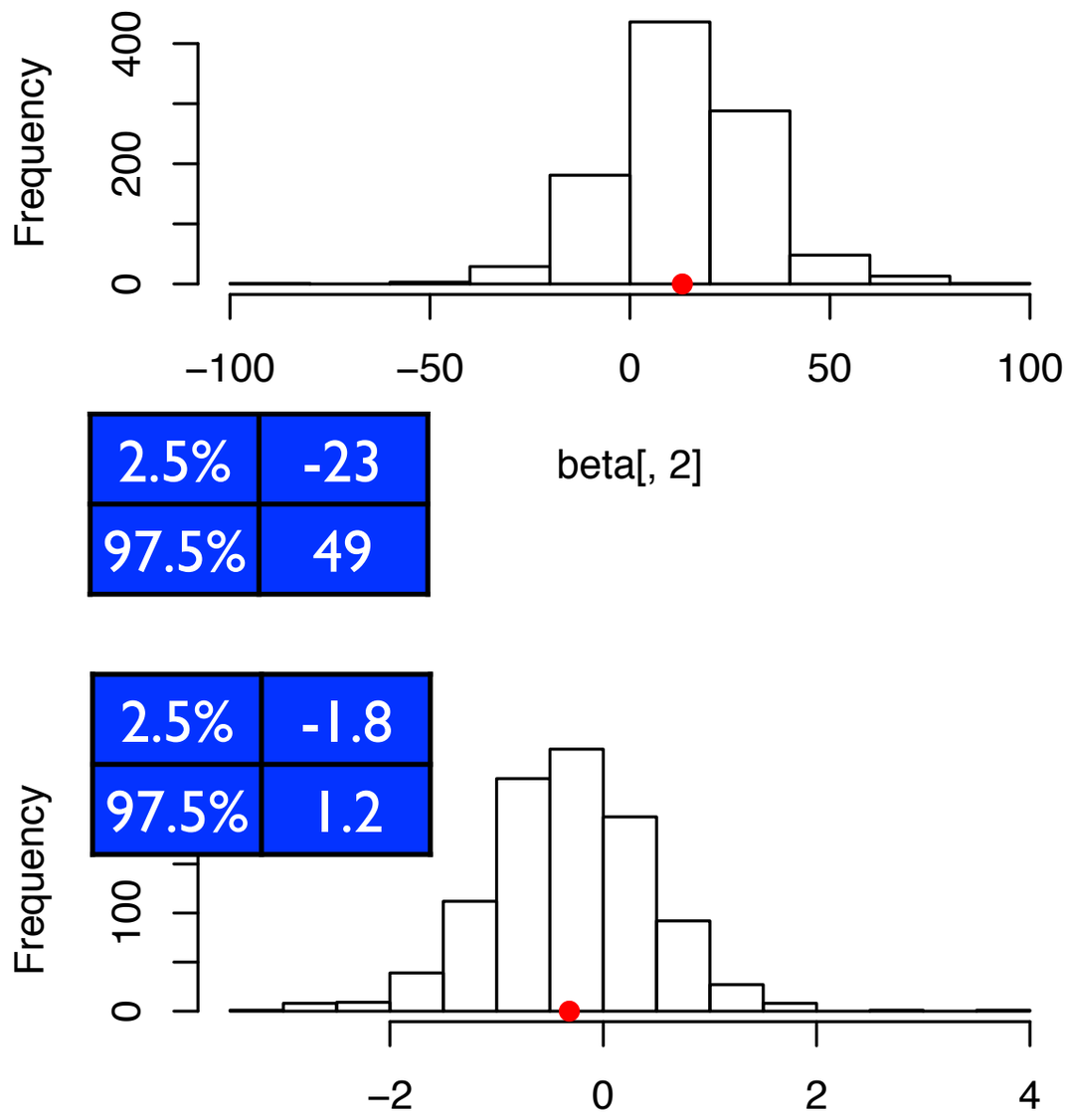
$$= \mathrm{IG}\left(\frac{\nu_0+n-p}{2}, \frac{\nu_0\sigma_0^2+n\hat{\sigma}^2}{2}\right)$$

Since we can sample from both of these distributions, samples from the joint posterior $p(\beta,\sigma^2|y,X)$ may be obtained by MC

# Example: Oxygen intake

We will use the independent Jeffreys' prior for this example with the MC procedure using the marginal posterior $p(\sigma^2|y, X)$



# Example: effect of aerobics?

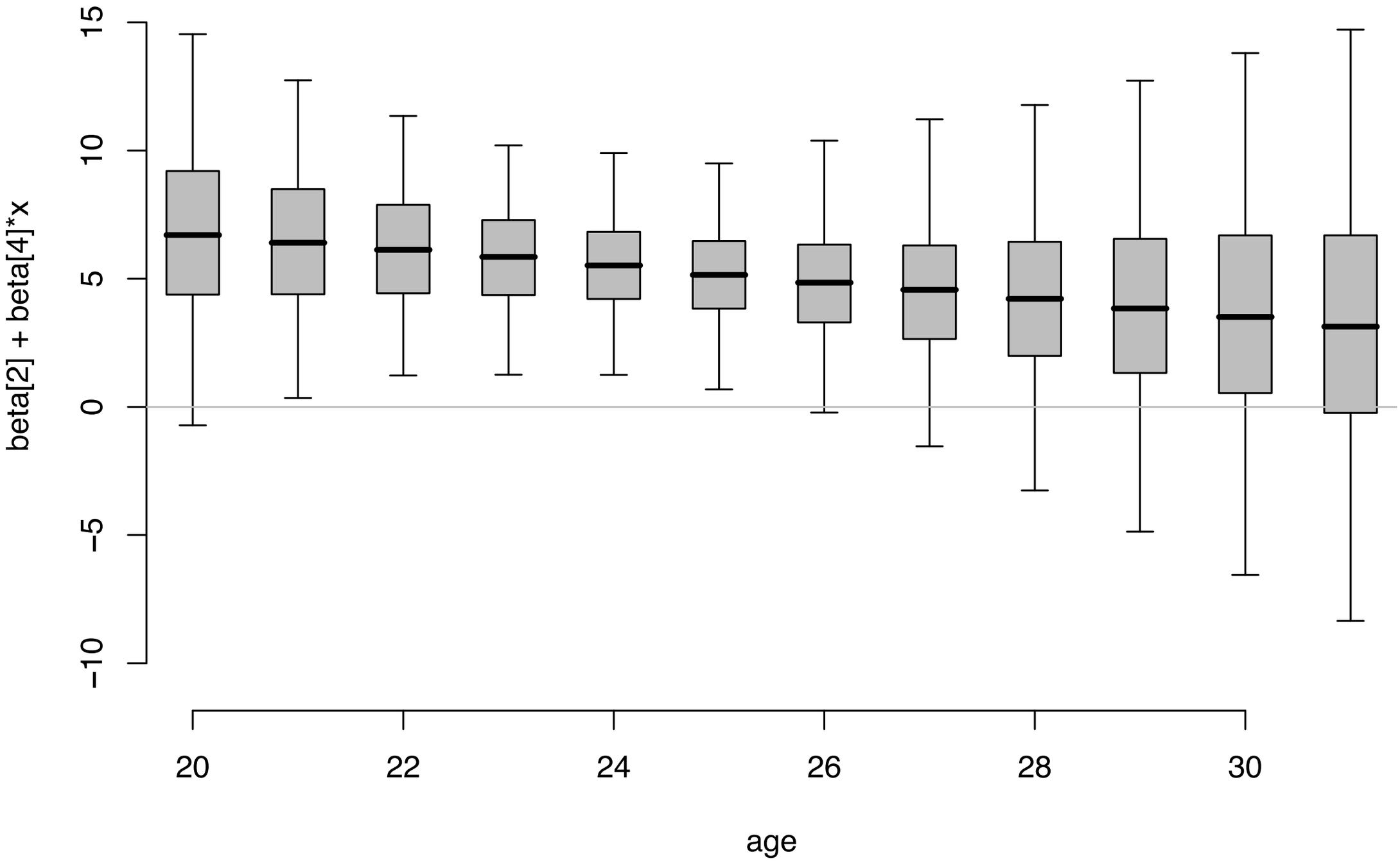
The posterior distributions seem to suggest only weak evidence of a difference between the two groups, as the 95% CIs for $\beta_2$ and $\beta_4$ both contain zero

However, these parameters themselves do not quite tell the whole story

According to the model, the average difference in $y$ between two people of the same age $x$ but in different exercise programs is $\beta_2 + \beta_4 x$

Thus, the posterior distribution for the effect of the aerobics program over the running program is obtained via the posterior distribution of $\beta_2 + \beta_4 x$ for each $x$

# Example: effect of aerobics depends on age




This indicates reasonably strong evidence of a difference at young ages and less at older ones

# Prediction (forecasting)

Suppose that we wish to obtain the posterior predictive distribution $Y(x^*)$ at a new location $x^*$

There are several ways in which we may go about sampling from this predictive distribution

The decomposition:

$$p(y^*|y, X, x^*) \overset{\text{(LTP)}}{=} \int p(y^*|x^*, \beta, \sigma^2) p(\beta, \sigma^2|y, X)\, d\beta\, d\sigma^2$$

(cond. prob.)

says that we may obtain MC samples as

$$y^{*(s)} \sim \mathcal{N}(x^{*\top}\beta^{(s)}, \sigma^{2(s)})$$

# Example: predictive quantiles