REGRESSION CASE STUDY

BST228

Speaker notes

No notes on this slide.

RECAP

- Developed the likelihood $p\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \tau\right)$ for linear regression.
- Derived conditional posterior distributions $\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau$ for regression coefficients and $\tau \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\beta}$.
- Considered limiting cases as sanity checks.
- Started applying methods to respiratory function dataset.

Speaker notes

- $\mathbf{y}$ are responses, $\mathbf{X}$ are features, $\boldsymbol{\beta}$ are regression coefficients, τ is observation noise precision.
- Linear regression likelihood can be expressed either in vector notation or as sum in log space for independent observations. The former is often more convenient for algebraic manipulation.
- We want to recover the prior mean $\boldsymbol{\nu}_0$ for large prior precision $\boldsymbol{\kappa}_0$ and the MLE for large observation precision τ.

TODAY

- Review derivations and limiting cases.
- Gibbs sampler for regression.
- Analyze posterior samples.
- Posterior of best fit vs posterior predictive distribution.

Speaker notes

No notes on this slide.

CONDITIONAL POSTERIOR FOR REGRESSION COEFFICIENTS β (1 / 2)

The conditional posterior for regression coefficients $\boldsymbol{\beta}$ is

$$
\begin{aligned}
p\left(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau\right) &= p\left(\mathbf{y} \mid \mathbf{X}, \tau, \boldsymbol{\beta}\right) p\left(\boldsymbol{\beta}\right) \\
&\propto \exp\left[-\frac{\tau}{2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\intercal}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) - \frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\nu}_0\right)^{\intercal}\boldsymbol{\kappa}_0\left(\boldsymbol{\beta} - \boldsymbol{\nu}_0\right)\right] \\
&\propto \exp\left[-\frac{\tau}{2}\left(\mathbf{y}^{\intercal}\mathbf{y} - \mathbf{y}^{\intercal}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^{\intercal}\mathbf{X}^{\intercal}\mathbf{y} + \boldsymbol{\beta}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\boldsymbol{\beta}\right) - \frac{1}{2}\left(\boldsymbol{\beta}^{\intercal}\boldsymbol{\kappa}_0\boldsymbol{\beta} - \boldsymbol{\beta}^{\intercal}\boldsymbol{\kappa}_0\boldsymbol{\nu}_0 - \boldsymbol{\nu}_0^{\intercal}\boldsymbol{\kappa}_0\boldsymbol{\beta} + \boldsymbol{\nu}_0^{\intercal}\boldsymbol{\kappa}_0\boldsymbol{\nu}_0\right)\right],
\end{aligned}
$$

where the second line follows by substitution of the likelihood and prior from the previous slide using the multivariate normal density from slide 11 of lecture 11. The third line follows by distributing the inner products. Collecting terms gives

$$
p\left(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau\right) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\beta}^{\intercal}\boldsymbol{\kappa}_n\boldsymbol{\beta} - \boldsymbol{\beta}^{\intercal}\left(\boldsymbol{\kappa}_0\boldsymbol{\nu}_0 + \tau\mathbf{X}^{\intercal}\mathbf{y}\right) - \left(\boldsymbol{\kappa}_0\boldsymbol{\nu}_0 + \tau\mathbf{X}^{\intercal}\mathbf{y}\right)^{\intercal}\boldsymbol{\beta}\right)\right],
$$

where we have defined $\boldsymbol{\kappa}_n = \left(\boldsymbol{\kappa}_0 + \tau\mathbf{X}^{\intercal}\mathbf{X}\right)$. This term looks just like the precision matrix of a multivariate normal distribution. On the next slide, we consider the linear terms in $\boldsymbol{\beta}$.

Speaker notes

No notes on this slide.

CONDITIONAL POSTERIOR FOR REGRESSION COEFFICIENTS β (2 / 2)

Without changing the result, we insert $\boldsymbol{\kappa}_n \boldsymbol{\kappa}_n^{-1} = \mathbf{I}$ between $\boldsymbol{\beta}$ and $(\boldsymbol{\kappa}_0 \boldsymbol{\nu}_0 + \tau \mathbf{X}^\intercal \mathbf{y})$ to get

$$
\begin{aligned}
p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau) &\propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\beta}^\intercal \boldsymbol{\kappa}_n \boldsymbol{\beta} - \boldsymbol{\beta}^\intercal \boldsymbol{\kappa}_n \boldsymbol{\kappa}_n^{-1}(\boldsymbol{\kappa}_0 \boldsymbol{\nu}_0 + \tau \mathbf{X}^\intercal \mathbf{y}) - (\boldsymbol{\kappa}_0 \boldsymbol{\nu}_0 + \tau \mathbf{X}^\intercal \mathbf{y})^\intercal \boldsymbol{\kappa}_n^{-1} \boldsymbol{\kappa}_n \boldsymbol{\beta}\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\beta}^\intercal \boldsymbol{\kappa}_n \boldsymbol{\beta} - \boldsymbol{\beta}^\intercal \boldsymbol{\kappa}_n \boldsymbol{\nu}_n - \boldsymbol{\nu}_n^\intercal \boldsymbol{\kappa}_n \boldsymbol{\beta}\right)\right],
\end{aligned}
$$

where we defined $\boldsymbol{\nu}_n = \boldsymbol{\kappa}_n^{-1}(\boldsymbol{\kappa}_0 \boldsymbol{\nu}_0 + \tau \mathbf{X}^\intercal \mathbf{y})$. We can now complete the square to obtain

$$
p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\nu}_n)^\intercal \boldsymbol{\kappa}_n (\boldsymbol{\beta} - \boldsymbol{\nu}_n)\right]
$$

and the conditional distribution is multivariate normal:

$$
\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau \sim \text{Normal}\left((\boldsymbol{\kappa}_0 + \tau \mathbf{X}^\intercal \mathbf{X})^{-1}(\boldsymbol{\kappa}_0 \boldsymbol{\nu}_0 + \tau \mathbf{X}^\intercal \mathbf{y}), (\boldsymbol{\kappa}_0 + \tau \mathbf{X}^\intercal \mathbf{X})^{-1}\right).
$$

Speaker notes

No notes on this slide.

CONDITIONAL POSTERIOR FOR OBSERVATION PRECISION τ

The conditional posterior for observation precision τ is

$$\begin{aligned} p\left(\tau \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\beta}\right) &= p\left(\mathbf{y} \mid \mathbf{X}, \tau, \boldsymbol{\beta}\right) p\left(\tau\right) \\ &\propto \tau^{n/2} \exp\left[-\frac{\tau}{2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\mathsf{T}}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right] \tau^{a_0 - 1} \exp\left[-b_0 \tau\right]. \end{aligned}$$

Collecting terms, we recognize the kernel of a gamma distribution with parameters

$$\begin{aligned} a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\mathsf{T}}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right). \end{aligned}$$

Speaker notes

No notes on this slide.

LIMITING CASES (1 / 2)

- For large prior precision $\boldsymbol{\kappa}_0$, we recover our prior best guess at the regression coefficients:

$$\begin{aligned}\lim_{\boldsymbol{\kappa}_0 \to \infty} \mathbb{E}\left[\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau\right] &= \lim_{\boldsymbol{\kappa}_0 \to \infty} \left(\boldsymbol{\kappa}_0 + \tau \mathbf{X}^\intercal \mathbf{X}\right)^{-1} \left(\boldsymbol{\kappa}_0 \boldsymbol{\nu}_0 + \tau \mathbf{X}^\intercal \mathbf{y}\right) \\ &= \lim_{\boldsymbol{\kappa}_0 \to \infty} \boldsymbol{\kappa}_0^{-1} \boldsymbol{\kappa}_0 \boldsymbol{\nu}_0 \\ &= \boldsymbol{\nu}_0.\end{aligned}$$

- For large observation precision τ, we recover the maximum likelihood estimate:

$$\begin{aligned}\lim_{\tau \to \infty} \mathbb{E}\left[\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau\right] &= \lim_{\tau \to \infty} \left(\boldsymbol{\kappa}_0 + \tau \mathbf{X}^\intercal \mathbf{X}\right)^{-1} \left(\boldsymbol{\kappa}_0 \boldsymbol{\nu}_0 + \tau \mathbf{X}^\intercal \mathbf{y}\right) \\ &= \lim_{\tau \to \infty} \tau^{-1} \left(\mathbf{X}^\intercal \mathbf{X}\right)^{-1} \tau \mathbf{X}^\intercal \mathbf{y} \\ &= \left(\mathbf{X}^\intercal \mathbf{X}\right)^{-1} \mathbf{X}^\intercal \mathbf{y},\end{aligned}$$

  where the second equality follows because $\left(a\mathbf{B}\right)^{-1} = a^{-1}\mathbf{B}^{-1}$ and the third because scalars commute with inner products.

Speaker notes

- We consider limiting cases as sanity checks for the derivation of the conditional distributions.

LIMITING CASES (2 / 2)

For the limit $n \to \infty$, we need to rearrange the expression for $\boldsymbol{\nu}_n$ slightly because it does not explicitly depend on n:

$$\boldsymbol{\nu}_n = \left(\boldsymbol{\kappa}_0 + \tau n \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_j^\intercal \right] \right)^{-1} \left(\boldsymbol{\kappa}_0 \boldsymbol{\nu}_0 + \tau n \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i y_i \right] \right).$$

In the limit, the expressions in brackets converge to expectations under an infinite population:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_j^\intercal = \mathbb{E}\left[\mathbf{x}\mathbf{x}^\intercal\right]$$

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i y_i = \mathbb{E}\left[\mathbf{x}y\right]$$

Substituting yields

$$\begin{aligned}
\lim_{n \to \infty} \boldsymbol{\nu}_n &= \lim_{n \to \infty} \left(\boldsymbol{\kappa}_0 + \tau n \mathbb{E}\left[\mathbf{x}\mathbf{x}\right] \right)^{-1} \left(\boldsymbol{\kappa}_0 \boldsymbol{\nu}_0 + \tau n \mathbb{E}\left[\mathbf{x}y\right] \right) \\
&= \lim_{n \to \infty} \left(\tau n \right)^{-1} \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}\right] \right)^{-1} \tau n \mathbb{E}\left[\mathbf{x}y\right] \\
&= \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}\right] \right)^{-1} \mathbb{E}\left[\mathbf{x}y\right].
\end{aligned}$$

The second and third equalities follow the same argument as for the limit $\tau \to \infty$ on the previous slide.

Speaker notes

- In a Bayesian setting, we rarely think about infinite populations except for limiting cases such as this one.

```
> # Load the data and construct features.
> data <- read.csv("../lecture_11_regression/ventilation_change.csv")
> y <- data$change
> Z <- cbind(data$group == "aerobic", data$age)
> X <- cbind(rep(1, length(y)), Z[, 1], Z[, 2], Z[, 1] * Z[, 2])
> n_features <- ncol(X)
> n_subjects <- nrow(X)
>
> # Hyperparameters.
> nu_0 <- rep(0, n_features)
> kappa_0 <- diag(n_features) * 1e-4
> a_0 <- 1e-3
> b_0 <- 1e-3
>
```

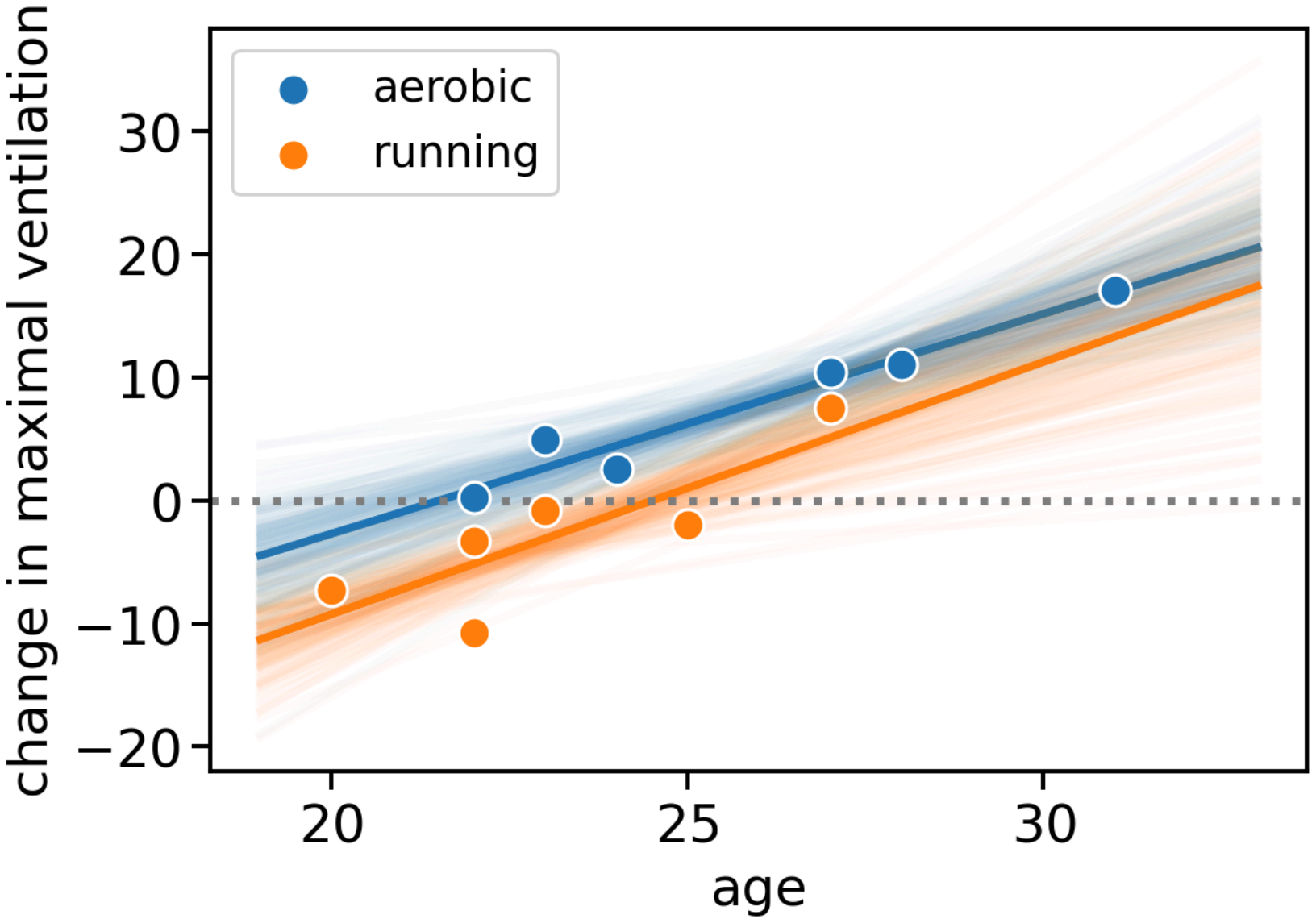
Speaker notes

- Lines #2-7 load the data `Z` and construct the design matrix `X`.
- #10-11 declare hyperparameters for a “non-informative” MVN prior for regression coefficients β.
- #12-13 declare hyperparameters for a “non-informative” gamma prior for observation precision τ.

```
> source("ventilation_init.R")
> source("rmvnorm.R")
>
> # Initialize the sampler.
> n_samples <- 10000
> beta <- rnorm(n_features)
> tau <- rgamma(1, 5, 5)
> samples <- list(beta = matrix(nrow = n_samples, ncol = n_features),
+                 tau = numeric(n_samples))
>
> set.seed(42)
> for (i in 1:n_samples) {
+   # Sample beta given tau.
+   # beta <- rmvnorm(1, ...)
+   # Sample tau given beta.
+   # tau <- rgamma(1, ...)
+   # Record values.
+   samples$beta[i, ] <- beta
+   samples$tau[i] <- tau
+ }
>
```

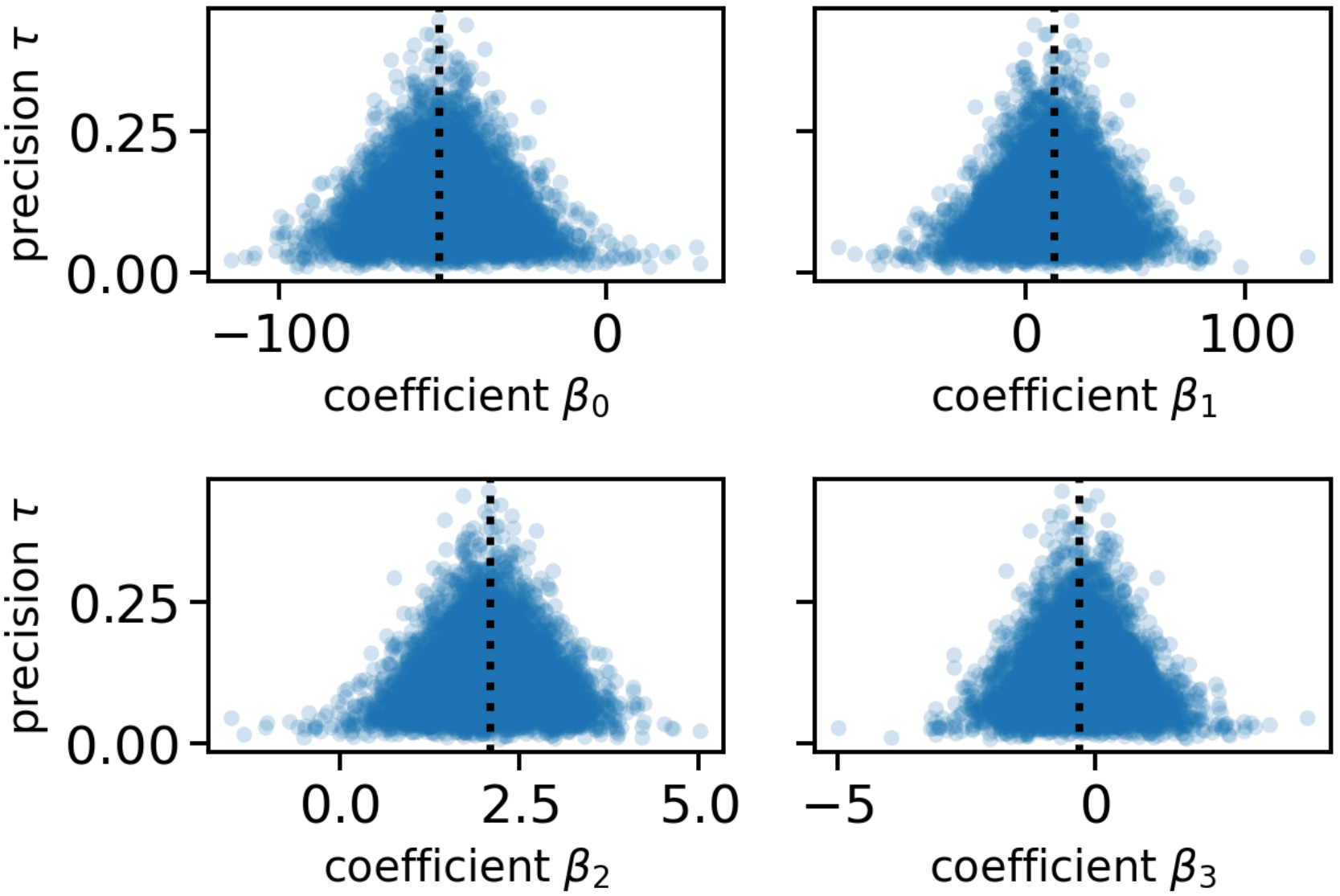
Speaker notes

- Line #1 loads from previous file.
- #2 loads helper `rmvnorm` to sample from MVN (not included in R).
- #5-9 set up variables to store samples.



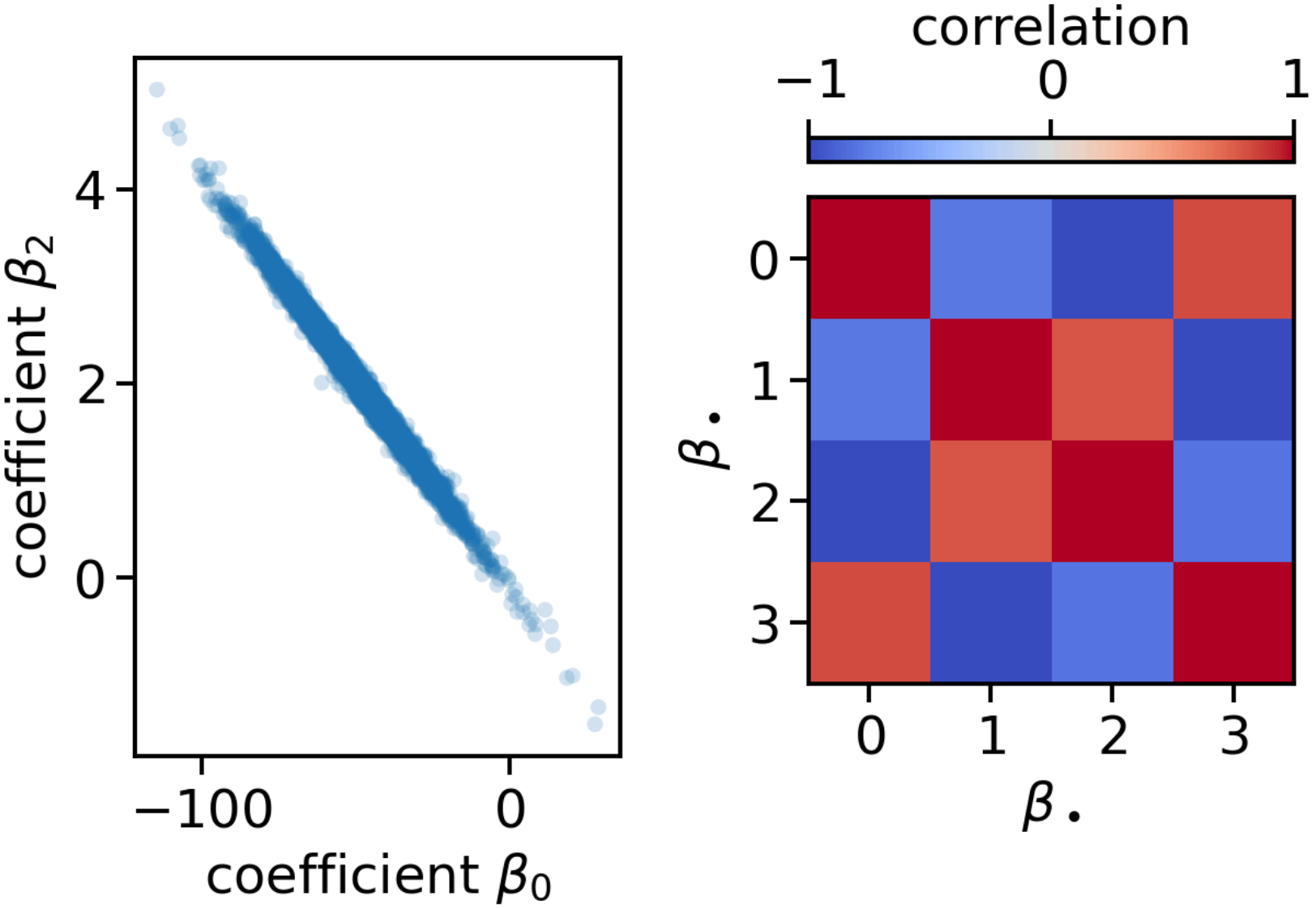
Speaker notes

- We obtain samples of the regression coefficients β and evaluate the *predictor* $\hat{\mathbf{y}} = \mathbf{X}\beta$ for different ages and exercise regimes.
- Transparent lines are individual samples and give us a notion of variability.
- Solid lines are the posterior mean of predictors and are our best-fit line.
- We have “succeeded” in the sense of having sensible fits with uncertainty quantification.
- We next analyze posterior samples to get a deeper understanding of regression, the shape of the posterior, implications for inference and prediction.



Speaker notes

- We have four *pair plots* of posterior samples of precision τ (y-axis) against each of the coefficients β.
- Pair plots are simple yet one of the most powerful tools for analyzing posterior distributions.
- Vertical lines are the maximum likelihood estimates.
- We have a distinctive “funnel” or “pyramid” shape in these plots.



Speaker notes

- Consider a pair plot of the intercept β_0 and age regression coefficient β_2. There is very strong correlation.
- Second panel is heatmap of posterior correlation matrix. Red is correlated, blue is anticorrelated.
- All parameters exhibit strong (anti-)correlation.

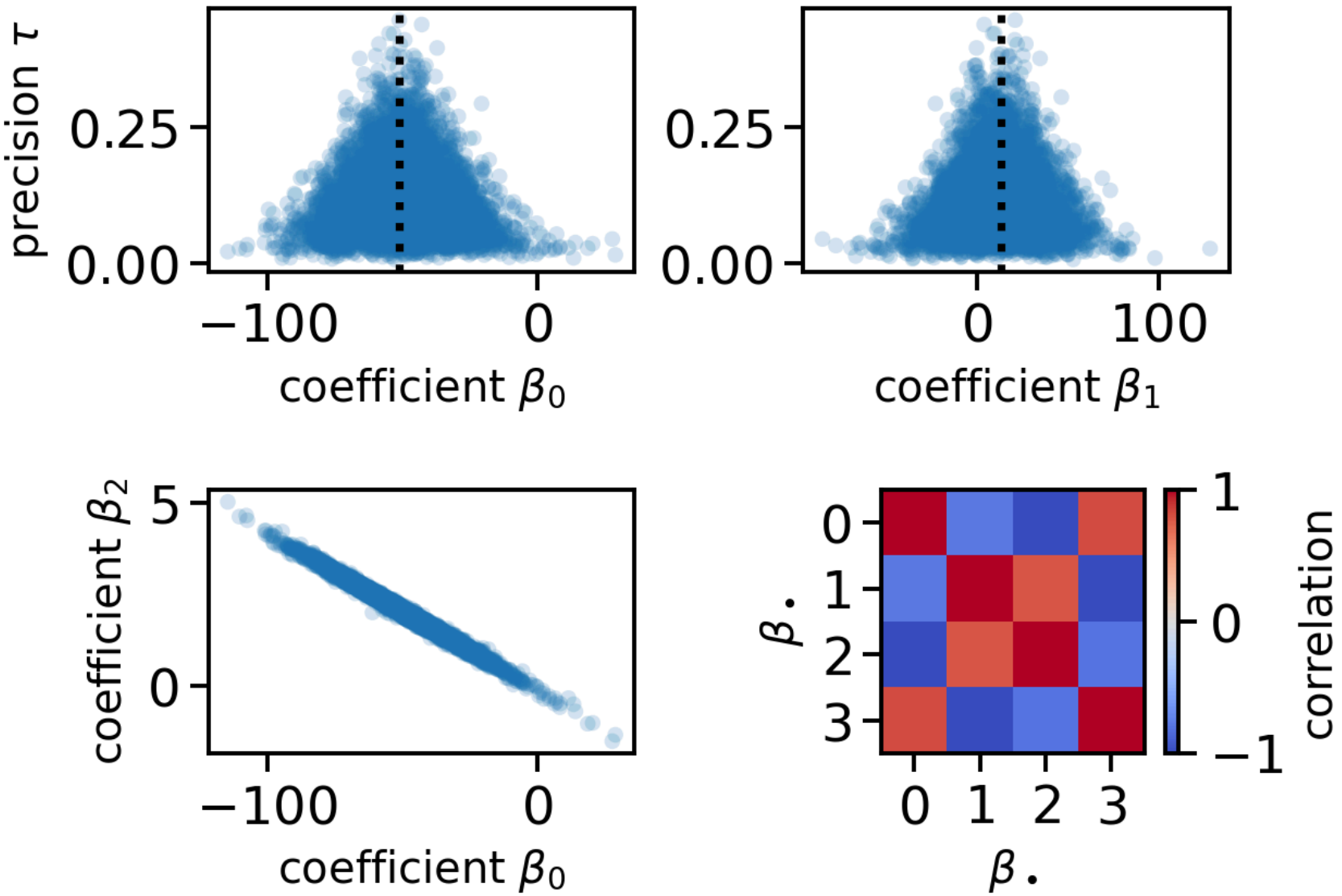
COLLABORATION

For 3 minutes, explore these questions by yourself. For 5 minutes, discuss with a group. Note your insights on post-its.

- What is the origin of the funnels in $\tau - \beta$ pair plots?
- Why are the regression coefficients correlated?
- What implications do your insights have for samplers?
- How does this affect your interpretation of regression coefficients?

Speaker notes

- Funnels are due to large precision giving rise to more precise parameter estimates.
- Regression coefficients are correlated due to co-linearity of features.
- Un-blocked Gibbs and Metropolis samplers will explore very slowly for correlated posteriors. Same for funnels because we need to explore the “neck” and “bulk” separately. Rejection samplers would have very high rejection rate.
- Interpretation of coefficients is challenging because marginal posteriors do not accurately reflect the joint posterior: We cannot think of parameters independently but they are inherently correlated.



Speaker notes

- Composite of previous two figures, omitting pair plots for β_2 and β_3.
- Figure serves as reference for students during collaborative task.

PARAMETER CORRELATION (1 / 2)

For negligible $\boldsymbol{\kappa}_0$,

$$\begin{aligned}\boldsymbol{\kappa}_n &= \tau \mathbf{X}^\intercal \mathbf{X} \\ &= \tau n \left(\bar{\mathbf{x}} \bar{\mathbf{x}}^\intercal + \boldsymbol{\Sigma} \right),\end{aligned}$$

where $\bar{x}_j = n^{-1} \sum_{i=1}^{n} X_{ij}$ is the sample mean and $\boldsymbol{\Sigma}_{jk} = n^{-1} \sum_{i=1}^{n} \left(X_{ij} - \bar{x}_j \right) \left(X_{ik} - \bar{x}_k \right)$ is the sample covariance.

Speaker notes

- We consider the $\boldsymbol{\kappa}_0 \ll \tau \mathbf{X}^\intercal \mathbf{X}$ for simplicity, but results hold more generally.
- The second equality reveals that the posterior precision comprises two terms: one depending on feature means $\bar{\mathbf{x}}$ and one on feature covariance $\boldsymbol{\Sigma}$.
- The mean-term gives rise to large correlation in the posterior distribution as we will see for a simple example on the next slide.

PARAMETER CORRELATION (1 / 2)

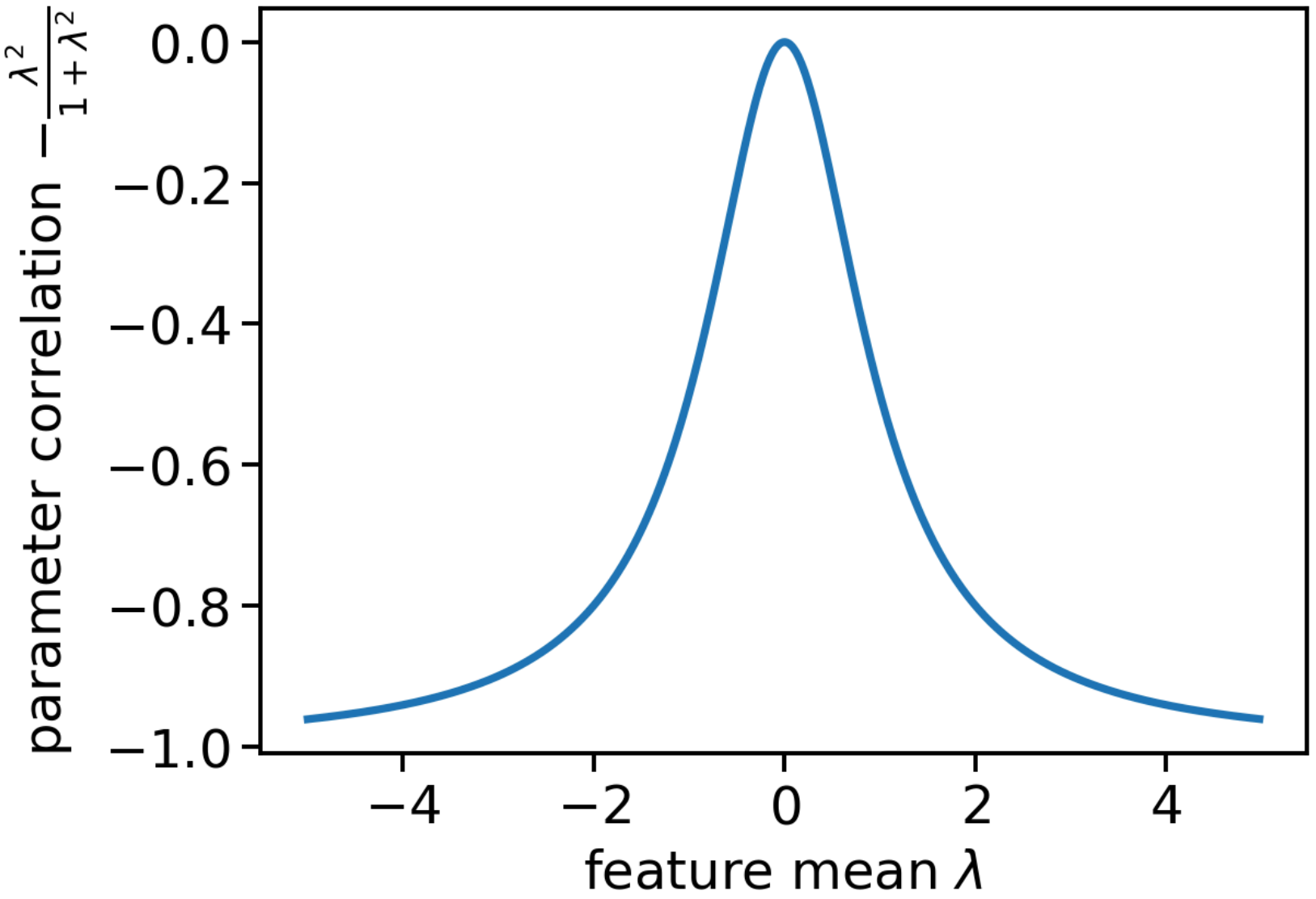
Suppose $\mathbf{x} \sim \textsf{Normal}\left((\lambda, \lambda)^\intercal, \mathbf{I}\right)$. Then

$$\boldsymbol{\kappa}_n = n\tau \begin{pmatrix} 1+\lambda^2 & \lambda^2 \\ \lambda^2 & 1+\lambda^2 \end{pmatrix}.$$

We next consider the correlation associated with the posterior covariance $\boldsymbol{\kappa}_n^{-1}$.

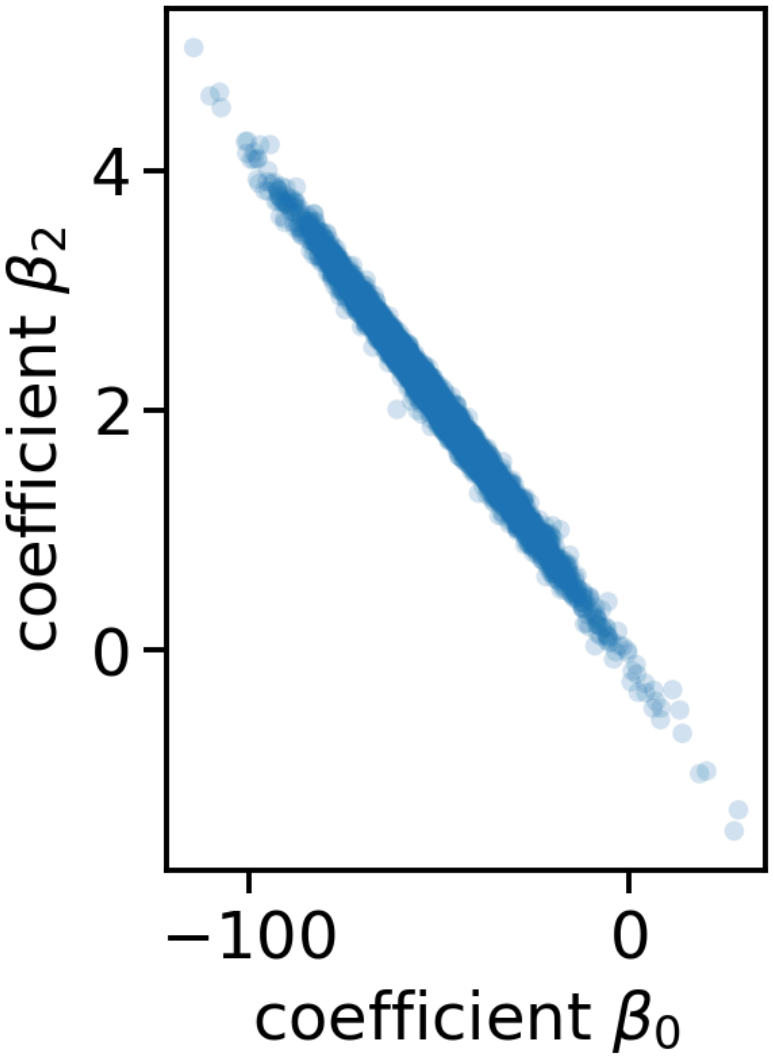
Speaker notes

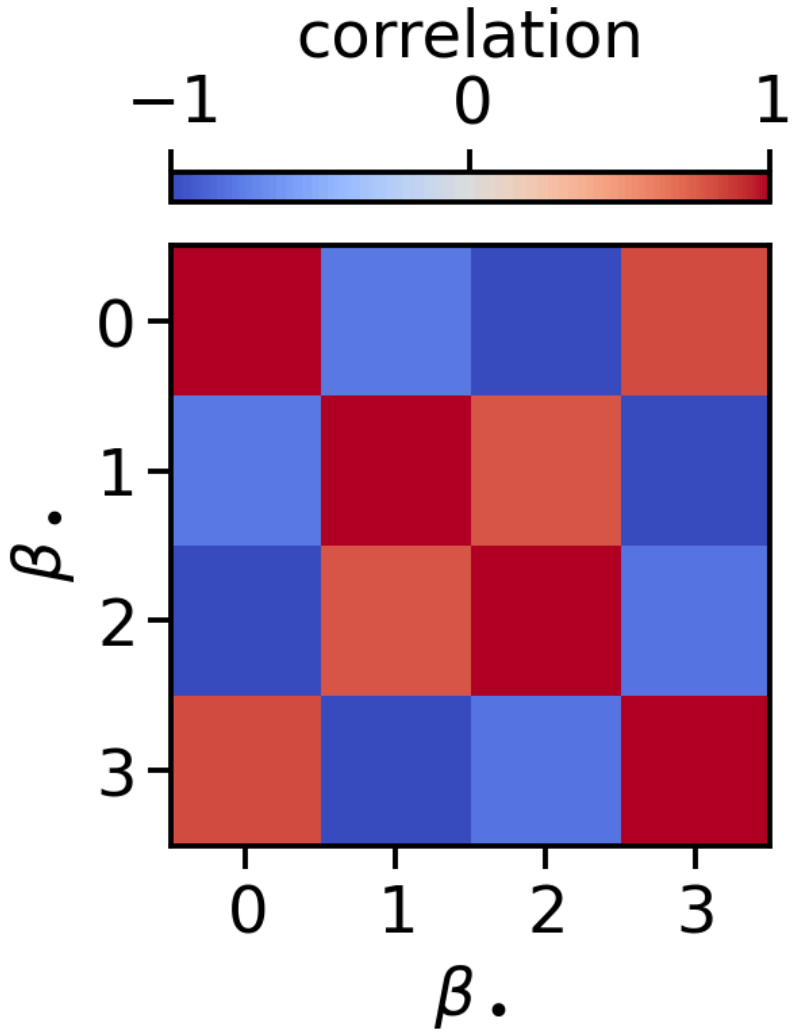
- In this hypothetical example, the independent bivariate features have mean λ and unit variance.
- The first line follows from substitution into the expression on the previous slide.
- Going through the algebra, we can show that the correlation corresponding to the precision $\boldsymbol{\kappa}_n$ is $-\frac{\lambda^2}{1+\lambda^2}$; exercise left to the reader.



Speaker notes

- Plotting the correlation coefficient between two parameters, we observe that the correlation is negative for non-zero mean λ.
- Why does this happen in the ventilation example? Because the features have non-zero mean: 1. has mean 1., 2., has mean of 0.5 (half in each treatment group), 3. has mean of average age, 4. has mean of `0.5 * (average age in aerobic group)`.
- This drives the strong anti-correlation among parameters and is an extreme case of co-linearity.
- Other correlations are due to the covariation between features.





Speaker notes

- Why might strong correlation be a problem?
- For Metropolis algorithms, we need a small proposal scale because we otherwise “step off” the high density region. This leads to slow exploration. Sampling regression coefficients iteratively using a Gibbs sampler also yields slow exploration. Rejection samplers have high rejection rate because the distribution is highly concentrated.
- Interpretation is challenging because we cannot consider properties of regression coefficients independently.

FUNNELS

Recall the Gibbs sampling steps for negligible prior parameters:

$$\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau \sim \text{Normal}\left(\left(\mathbf{X}^\intercal \mathbf{X}\right)^{-1} \mathbf{X}^\intercal \mathbf{y}, \tau^{-1} \left(\mathbf{X}^\intercal \mathbf{X}\right)^{-1} \right)$$

$$\tau \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\beta} \sim \text{Gamma}\left(\frac{n}{2}, \frac{\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^\intercal \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)}{2} \right).$$

Speaker notes

- By “negligible prior parameters” we mean that a_0, b_0, and $\boldsymbol{\kappa}_0$ are very small.
- Large τ implies samples of $\boldsymbol{\theta}$ very close to the MLE. This implies large τ because the residuals are small. This implies $\boldsymbol{\beta}$ close to the MLE, …
- The inverse logic applies if we start with small τ.
- Walking from the “neck” of the funnel to the “bulk” of the funnel can take a long time.
- These funnels appear a lot in hierarchical models.

NON-TRIVIAL GEOMETRY IS EVIL!

Speaker notes

- Non-trivial geometry messes with all practical samplers. As a Bayesian biostatistician, you will likely spend a lot of your time hunting down non-trivial geometry in your posterior.
- This may appear obscure and technical, but understanding these pathologies is essential for both interpreting posteriors and building models that can actually be fit in practice.
- We will explore funnels in more depth for hierarchical models and focus on reducing correlation for the rest of the lecture.

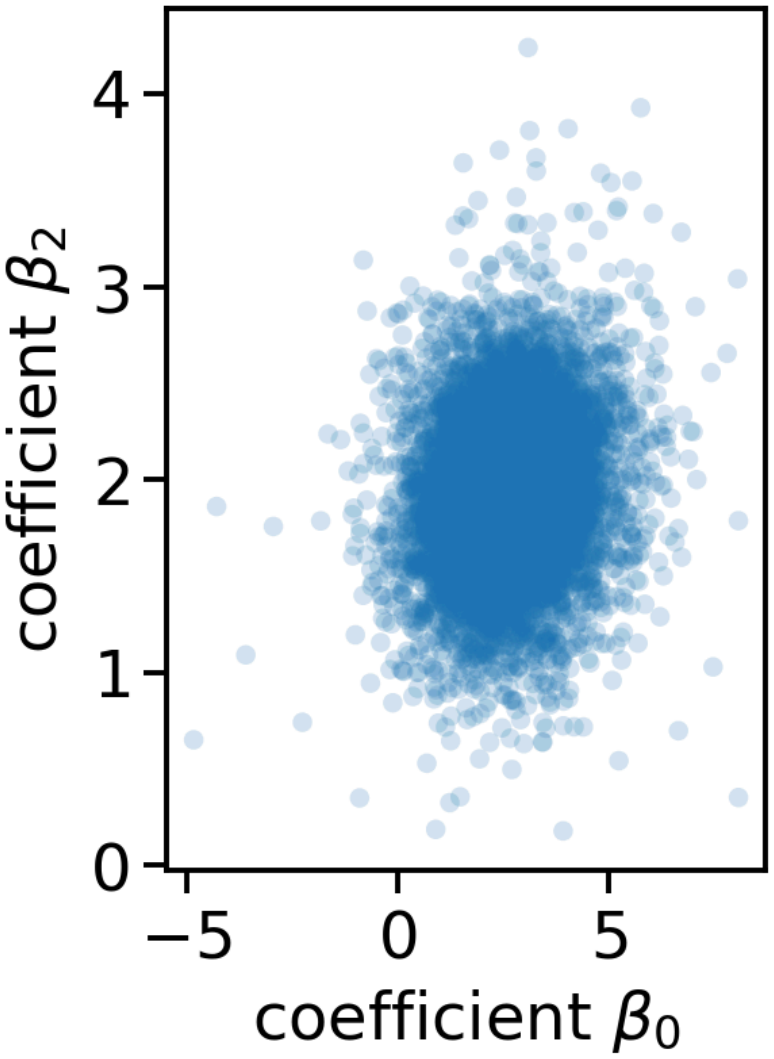
DE-MEANING FEATURES

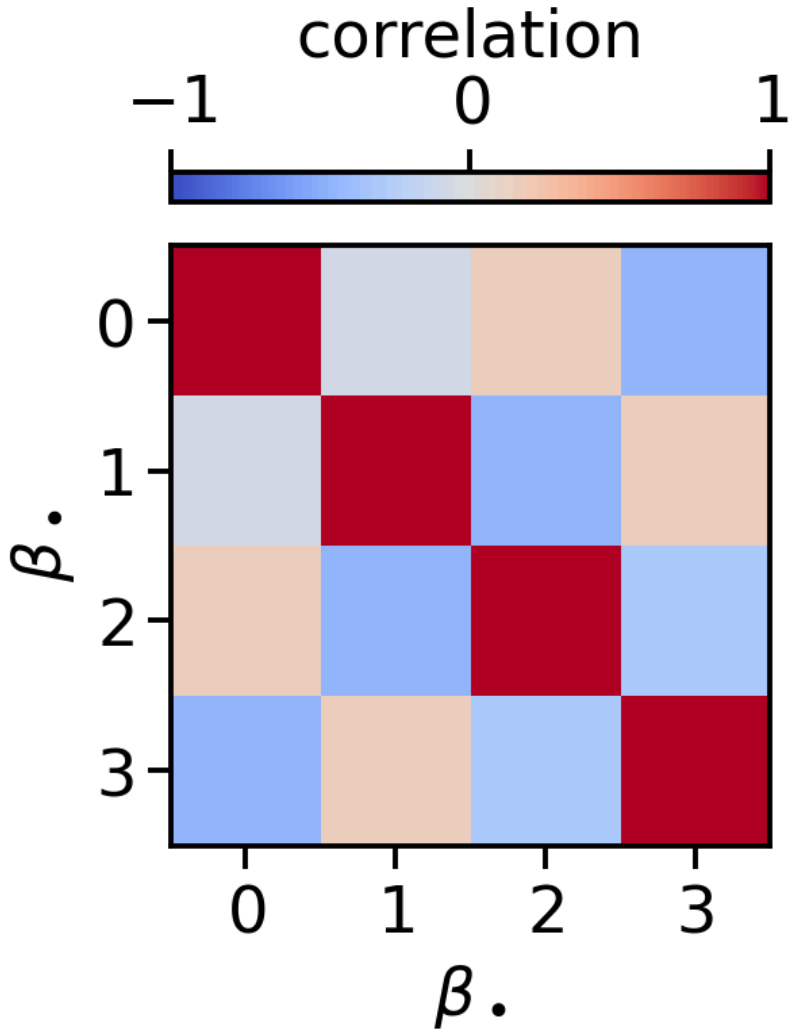
To attenuate posterior correlation, we define de-meaned features:

- $x_{i0} = 1$ is the intercept,
- $x_{i1} = z_{i1} - n^{-1} \sum_{i=1}^{n} z_{i1}$ is the de-meaned `aerobic` indicator,
- $x_{i2} = z_{i2} - n^{-1} \sum_{i=1}^{n} z_{i2}$ is the de-meaned age,
- $x_{i3} = x_{i1} \times x_{i2}$ is an interaction between de-meaned features.

Speaker notes

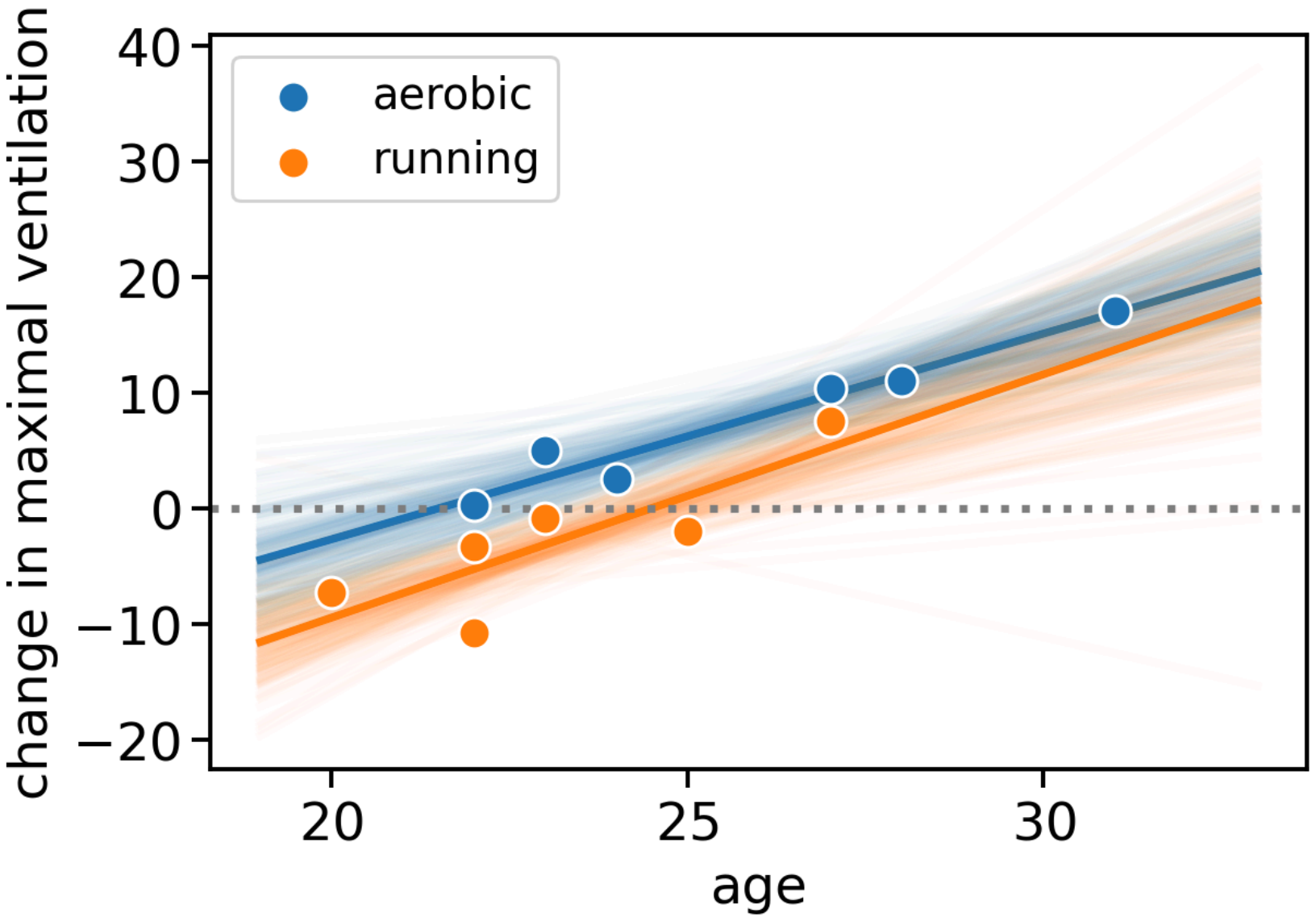
- These de-meaned features are an attempt to elminiate the correlation introduced by $\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathsf{T}}$ in the expression for the posterior precision $\boldsymbol{\kappa}_n$.





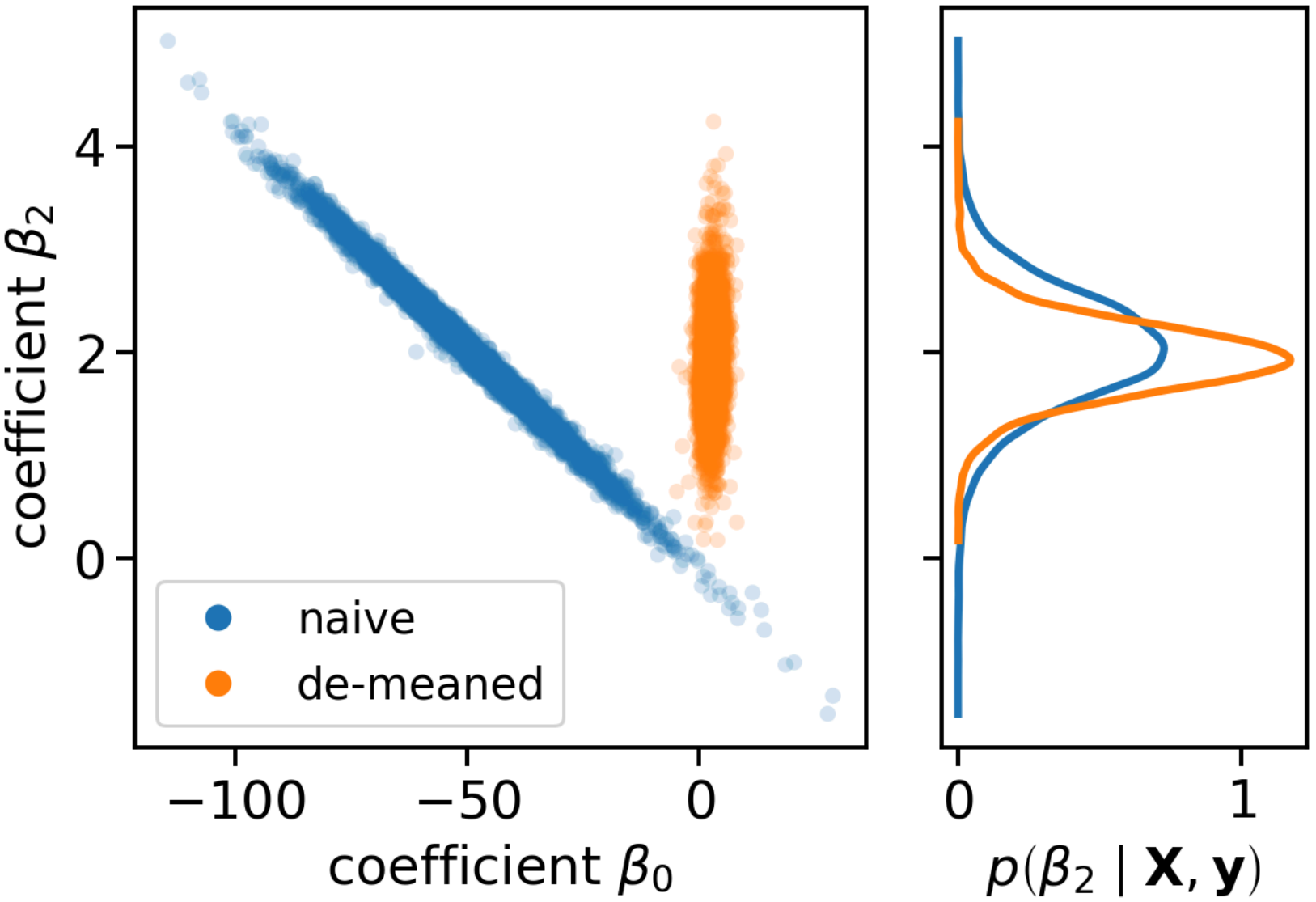
Speaker notes

- Re-running the sampler gives us both much more “circular” pair plots and a much less correlated posterior–exactly what we wanted.
- Aside: Reducing the complexity of the geometry is the reason ML books call for standardizing features. Optimization, just like sampling, is hard when the geometry is weird.



Speaker notes

- Let's look at the fit. The predictors $\mathbf{X}\beta$ are unchanged up to noise from the Gibbs sampler.
- This is a *necessity*: the effect of exercise on respiratory function cannot possibly depend on whether we subtract a constant from our features.



Speaker notes

- Left panel shows pair plots for the intercept and age coefficients for *naive* and *de-meaned* features.
- Posterior is uncorrelated and marginal uncertainty of coefficients is reduced.
- May be surprising but interpreting parameters is inherently difficult. They often do not have direct real-world interpretation.
- Distinction between interpreting changes in *predictions* (by construction, have real-world relevance) and changes in *parameters* ("just" part of the model) is particularly challenging for regression because they are so intimately coupled.
- Moral of the story: Making statements about predictions is (almost) always meaningful. Statements about model parameters may or may not have real-world relevance.

POSTERIOR PREDICTIVE DISTRIBUTION (PPD)

The PPD for new responses $\tilde{\mathbf{y}}$ given new features $\tilde{\mathbf{X}}$ is

$$p\left(\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}}, \mathbf{y}, \mathbf{X}\right) = \int d\boldsymbol{\beta}\, d\tau\, p\left(\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}}, \boldsymbol{\beta}, \tau\right) p\left(\boldsymbol{\beta}, \tau \mid \mathbf{X}, \mathbf{y}\right)$$

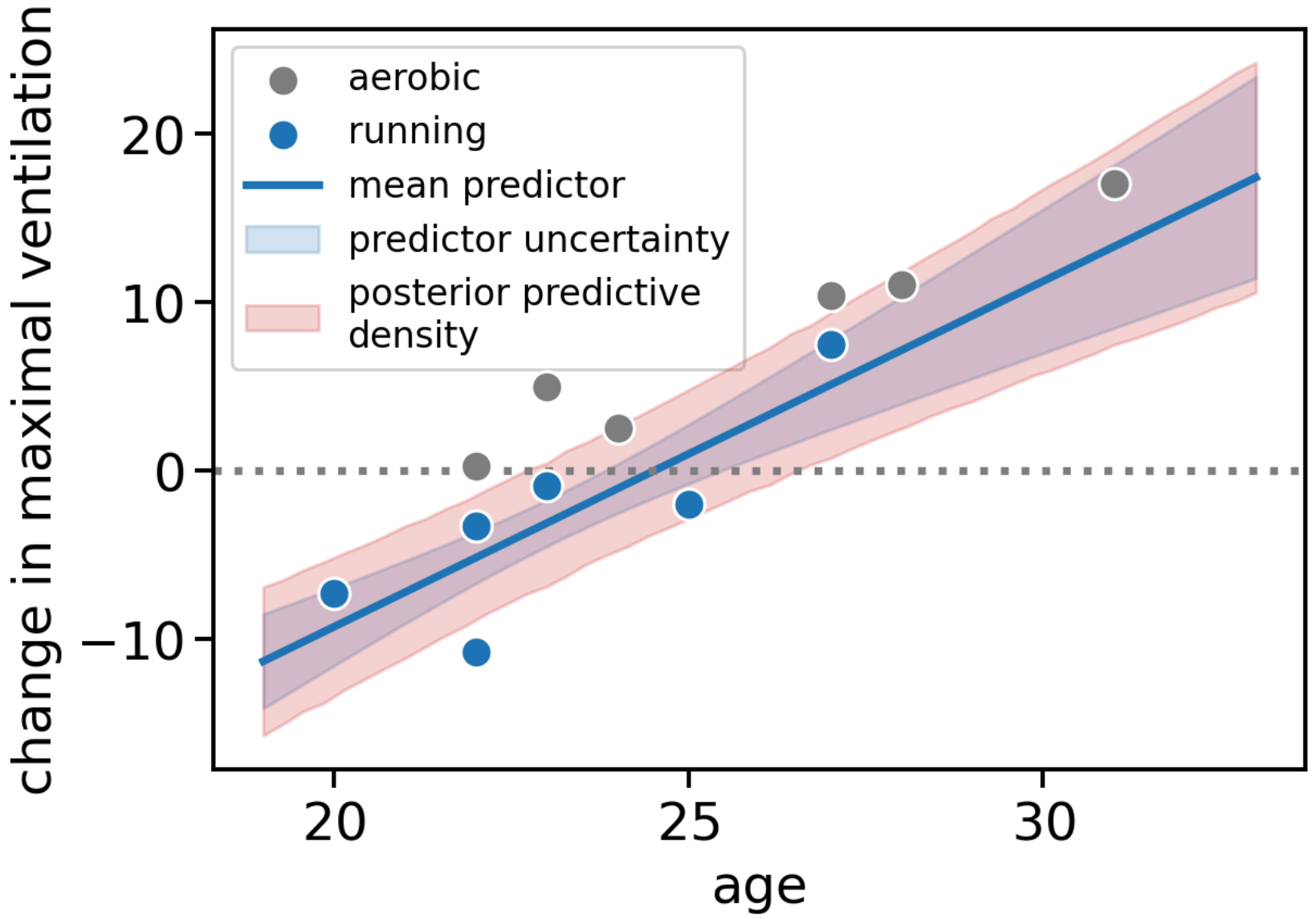
which simplifies to a Student's t-distribution.

We take a sampling approach and add the following to the Gibbs sampler:

$$\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}}, \boldsymbol{\beta}, \tau \sim \text{Normal}\left(\tilde{\mathbf{X}}\boldsymbol{\beta}, \tau^{-1}\mathbf{I}\right).$$

Speaker notes

- Given that predictions are the relevant quantity to focus on, let's consider the posterior predictive distribution, i.e., density of new responses $\tilde{\mathbf{y}}$ given new features $\tilde{\mathbf{X}}$ and data from our study $(\mathbf{X}, \mathbf{y})$.
- The first equality follows from the law of total probability. We forego the algebra required to obtain the Student's t-distribution (see lecture 7 for the derivation in the univariate case).
- Adding a sample statement to the Gibbs loop achieves the same goal yet with much less effort.
- We can treat the response $\tilde{\mathbf{y}}$ given new data $\tilde{\mathbf{X}}$ like any other variable in the model.



Speaker notes

- We focus on only the `running` subset to avoid clutter.
- The data are blue markers. The posterior mean of predictions is a blue line. Uncertainty (standard deviation band) *in the predictor* is the blue shaded area.
- Standard deviation in the predictive distribution is the red shaded area. Necessarily, the width of the posterior predictive is greater than the uncertainty in the posterior predictor because, in addition to uncertainty, it includes sampling noise for future observations.

NEXT

- Heteroskedastic regression.
- Generalized linear models.
- Stan.

Speaker notes

No notes on this slide.