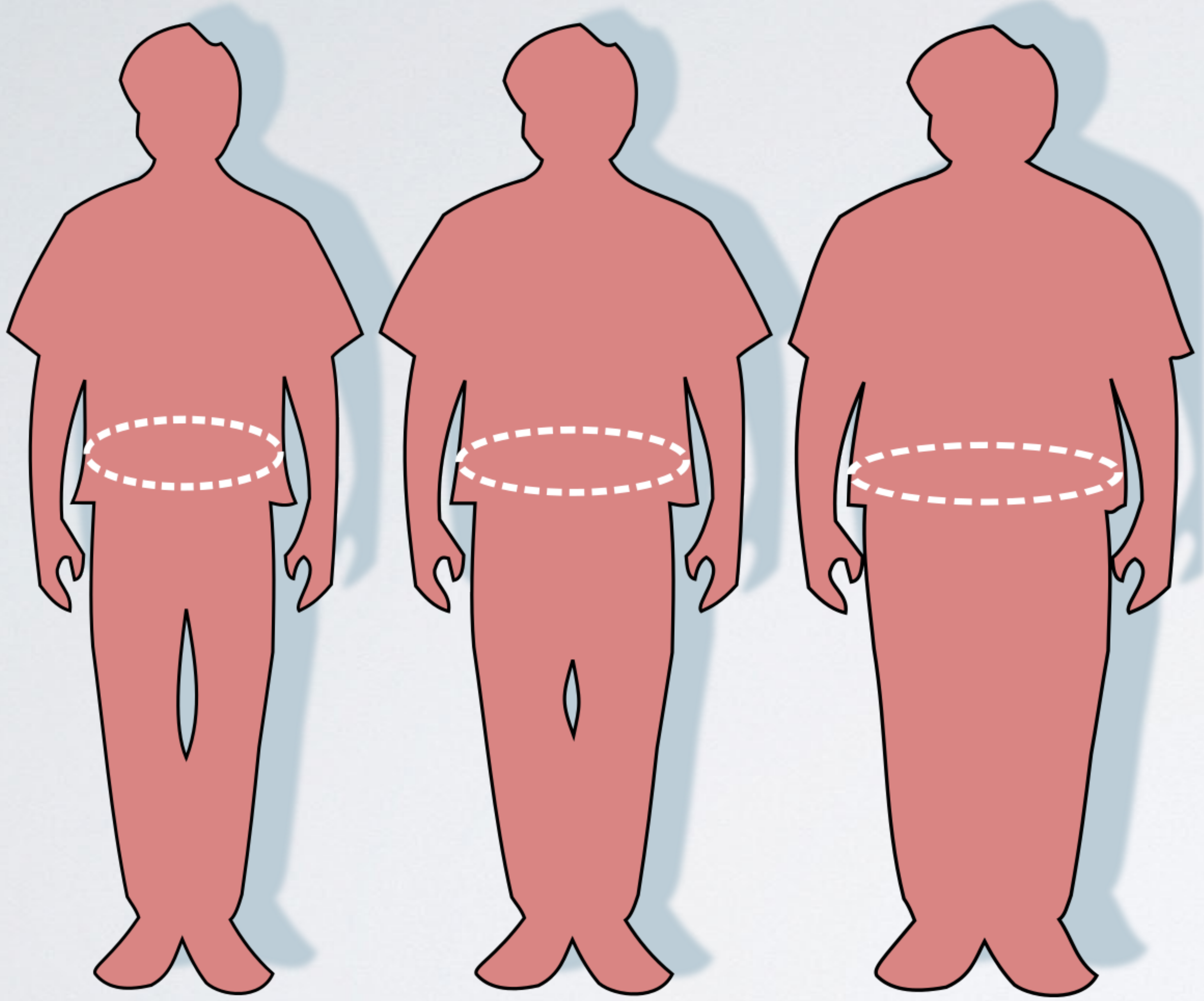
Bayesian linear regression

Dr. Merlise Clyde

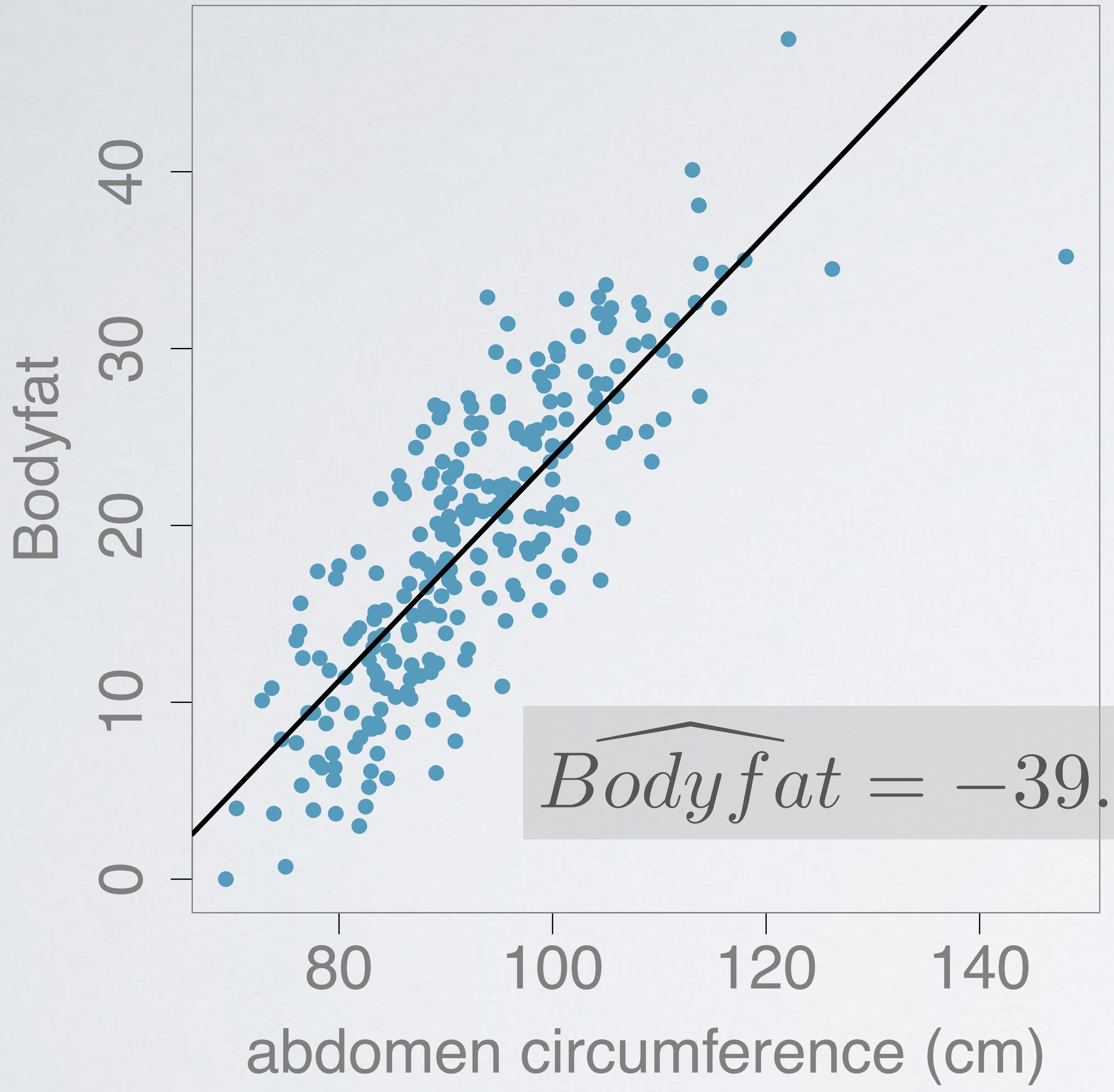
Duke University

body fat



Source: https://commons.wikimedia.org/wiki/File:Obesity-waist_circumference.svg

body fat data



$$\widehat{Bodyfat} = -39.28 + 0.63\,Abdomen$$

fitted values $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$

$\hat{\sigma}^2$ $\text{MSE} = \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^2}{n-2}$

model and prior

- model

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$$

- conjugate bivariate normal-gamma distribution

$$\alpha \mid \sigma^2 \sim \mathsf{N}(a_0, \sigma^2 S_\alpha)$$

$$\mathsf{cov}(\alpha, \beta \mid \sigma^2) = \sigma^2 S_{\alpha,\beta}$$

$$\beta \mid \sigma^2 \sim \mathsf{N}(b_0, \sigma^2 S_\beta)$$

$$1/\sigma^2 \sim \mathsf{G}(\nu_0/2, \nu_0\sigma_0^2/2)$$

reference prior and posterior distributions

reference prior

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

reference posterior

$$\beta \mid y_1, \ldots, y_n \sim t_{n-2}\left(\hat{\beta}, \mathrm{sd}(\beta)^2\right)$$

$$\alpha \mid y_1, \ldots y_n \sim t_{n-2}\left(\hat{\alpha}, \mathrm{sd}(\alpha)^2\right)$$

$$\alpha + \beta x_i \mid y_1, \ldots y_n \sim t_{n-2}\left(\hat{\alpha} + \hat{\beta} x_i, s^2_{y_i}\right)$$

$$s^2_{y_i} = s^2_{Y|X}\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)$$

estimates

	post. mean	post. sd	2.5%	97.5%
(intercept)	-39.28	2.66	-44.52	-34.04
abdomen	0.63	0.03	0.58	0.69

posterior mean $\pm\, t_{1-\alpha/2,n-2}$ posterior standard deviation

predicting body fat

- posterior predictive distribution for a new case

$$y_{n+1} = \alpha + \beta x_{n+1} + \varepsilon_{n+1}$$

- is also a Student t distribution with $n - 2$ df

$$y_{n+1} | y_1, \ldots y_n \sim t_{n-2}\left(\hat{y}_{n+1}, s^2_{y_{n+1}}\right)$$

$$\hat{y}_{n+1} = \hat{\alpha} + \hat{\beta} x_{n+1}$$

$$s^2_{y_{n+1}} = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)$$

predicting body fat (continued)

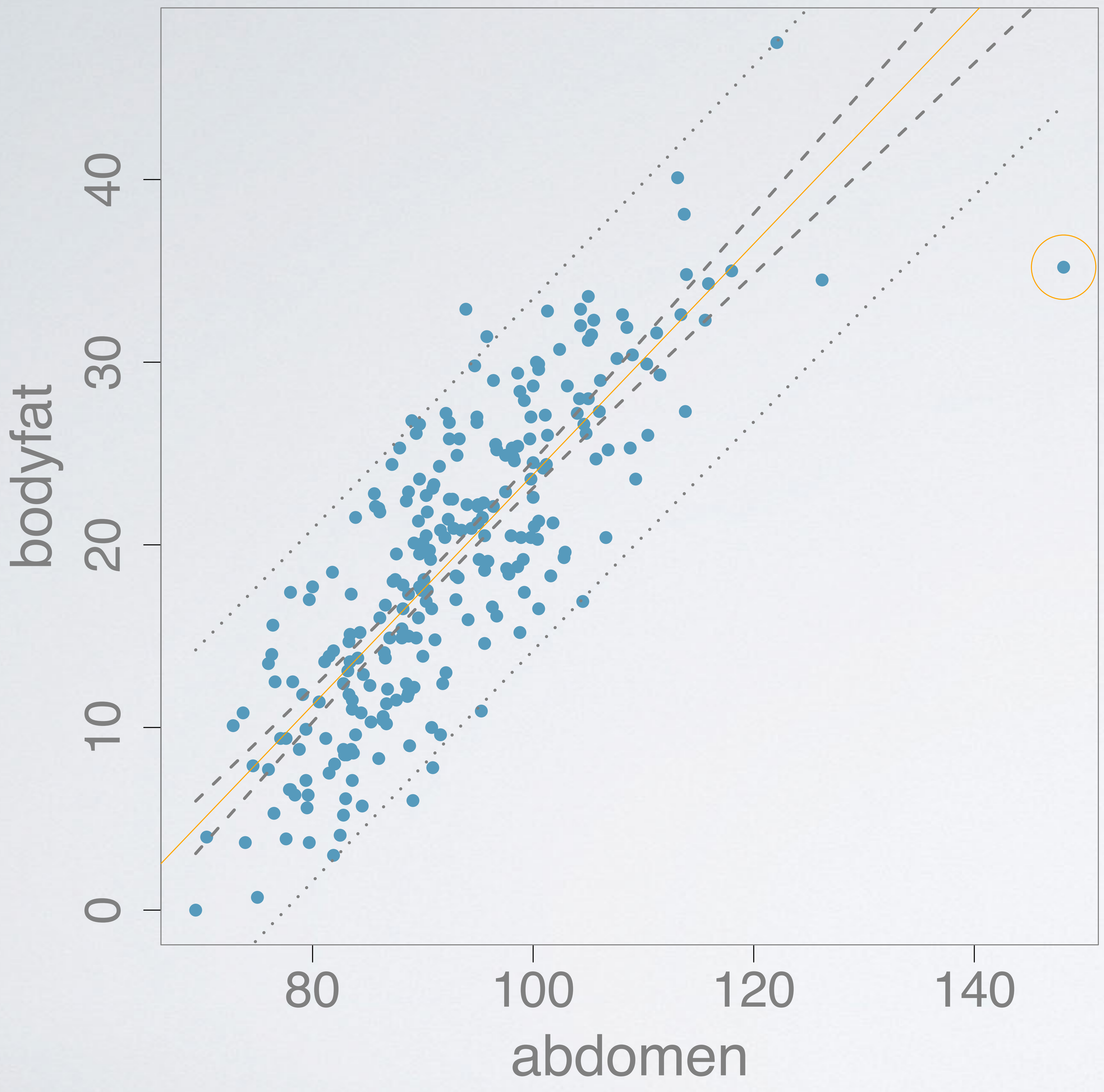
$$s^2_{y_{n+1}} = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)$$

posterior uncertainty about $\alpha + \beta x_{n+1}$

- depends on x_{n+1} spread
- is higher for x_{n+1} far from $\bar{x}$

additional variability $+s^2_{Y|X}$ due to ε_{n+1}

prediction intervals



summary

- under reference prior, point estimates and Bayesian credible intervals are equivalent to frequentist estimates and confidence intervals
- use standard software to obtain
- change in interpretation
- reference analysis