Bayesian Linear Regression

June 6, 2021

Why Bayesian Linear Regression?

Why Bayes? (Linear Regression Version)

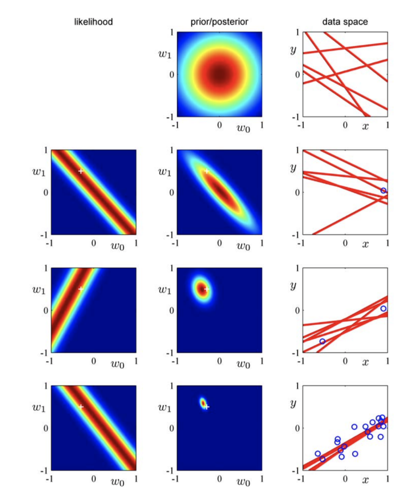
- Provides a *distribution* over regression lines
- Automatically supports model selection / complexity control.
- Easy access to nuanced inferential quantities.

Why Bayesian Linear Regression?

Distribution over regression lines

Bayesian Linear Regression

We learn a *distribution* over regression lines.



Sequential Bayesian learning for a simple linear model.

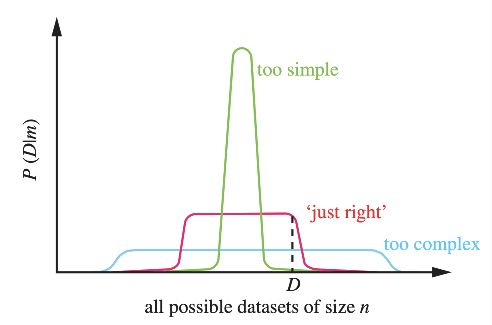
Image Credit: Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Why Bayesian Linear Regression?

Automatic model selection

Bayesian Occam's Razor

Remember this slide?

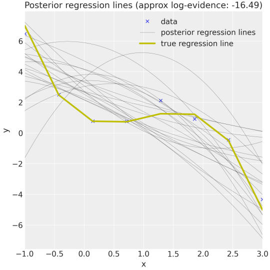


A *complex* model (shown in blue) spreads its mass over many more possible datasets, whereas a *simple* model (shown in green) concentrates its mass on a smaller fraction of possible data. Because probabilities have to sum to one, the complex model spreads its mass at the cost of not being able to model simple datasets as well as a simple model—this normalization is what results in an automatic Occam razor. Given any particular dataset, here indicated by the dotted line, we can use the marginal likelihood to reject both overly simple models, and overly complex models.

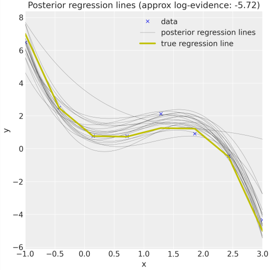
Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1984), 20110553.

Bayesian Occam's Razor

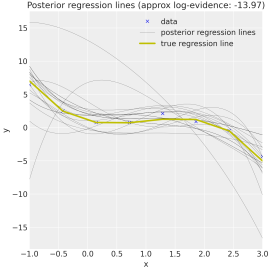
I generated $n = 8$ data points from a cubic distribution and used `NUTS` to fit Bayesian polynomials of various orders p.



Quadratic model ($p = 2$)



Cubic model ($p = 3$)

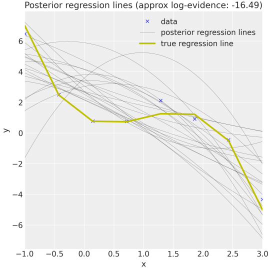


Quartic model ($p = 4$)

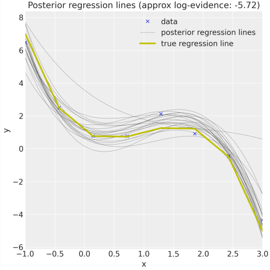
Observations

Bayesian Occam's Razor

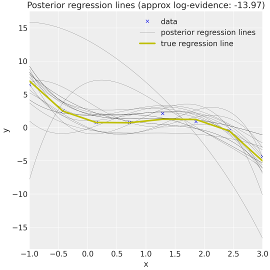
I generated $n = 8$ data points from a cubic distribution and used `NUTS` to fit Bayesian polynomials of various orders p.



Quadratic model ($p = 2$)



Cubic model ($p = 3$)



Quartic model ($p = 4$)

Observations

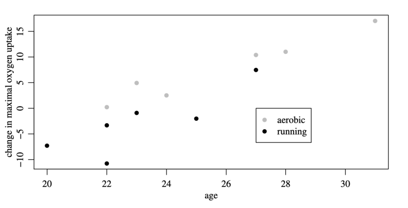
- Bayesian model selection works well here! The true (cubic) model has the highest evidence. The evidence is lower for models that are underfit (quadratic) or overfit (quartic).
- Posterior draws from the cubic model best match the true data generating process.
- Maximum likelihood doesn't do this. ML says: the higher the order, the *better* the fit.

Why Bayesian Linear Regression?

Easy access to nuanced inferential quantities

Exercise Program Dataset

(Hoff, A first course in Bayesian Statistical Methods)



Change in maximal O_2 uptake as a function of age and exercise program

Model

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

$$x_{i,1} = 1 \text{ for each subject } i$$

$$x_{i,2} = 0 \text{ if subject } i \text{ is on the running program, 1 if on aerobic}$$

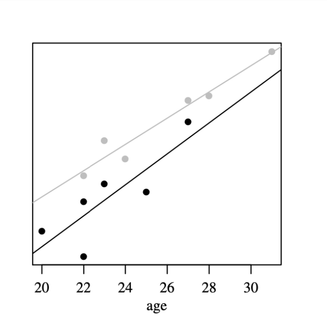
$$x_{i,2} = \text{age of subject } i$$

$$x_{i,4} = x_{i,2} \times x_{i,3}$$

Under this model, the conditional expectations for Y are:

$$\mathbb{E}[Y|\boldsymbol{x}] = \beta_1 + \beta_3 \times \text{age if } x_1 = 0, \text{ and}$$

$$\mathbb{E}[Y|\boldsymbol{x}] = (\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times \text{age if } x_1 = 1$$



Maximum likelihood regression lines for the O_2 uptake data.

$$\widehat{\beta}_{\text{ML}} = (-51.29, 13.11, 2.09, -.32)^T$$
$$\text{SE}(\widehat{\beta}_{\text{ML}}) = (12.25, 15.76, 0.53, 0.65)^T$$

Comparing the values of $\widehat{\beta}_{\text{ML}}$ to their standard errors suggests the evidence for differences between exercise programs is not very strong.

Bayesian inference agrees with the ML estimate, showing only weak evidence of a difference between exercise programs.

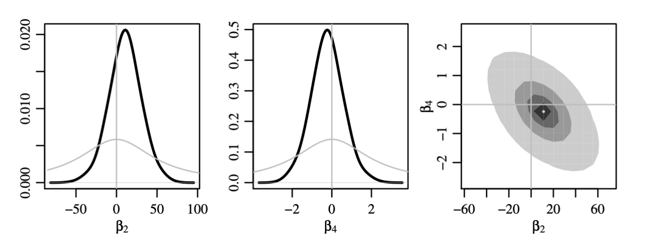


Figure 1: Posterior distributions for β_2 and β_4. The first two plots show the marginal prior distributions (grey) for comparison. The 95% posterior intervals for β_2 and β_4 both contain 0.

But the parameters by themselves don't tell the whole story.

We can also look at the posterior distributions of $\beta_2 + \beta_4 x$ for each age x.

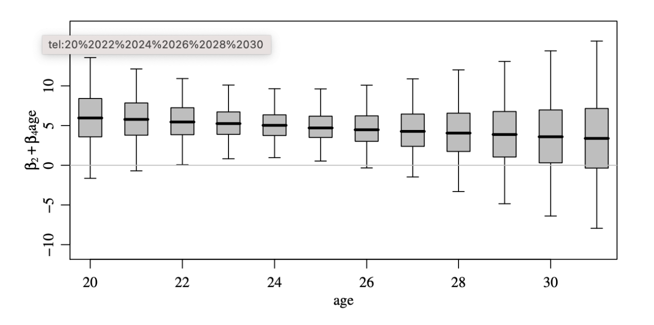


Figure 2: 95% confidence intervals for the difference in expected change scores between aerobics subjects and running subjects

This suggests reasonably strong evidence of a difference at young ages, and less evidence at older ones.

The model

Inference