STAT 153 - Introduction to Time Series Lecture Eleven

Fall 2022, UC Berkeley

Aditya Guntuboyina

September 30, 2022

1 Model Selection in Linear Regression

In the last lecture, we considered a linear regression model M of the form:

$$Y = X\beta + Z \qquad \text{with } Z_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2) \tag{1}$$

These models include time series models where we fit trend functions via polynomials and sinusoids at prefixed frequencies. For Bayesian model selection, we calculate the evidence of this model by integrating the likelihood multiplied by a prior on the parmaeters $\theta = (\beta, \sigma)$. In the last lecture, we saw the following approximate formula for the evidence:

$$\text{Evidence}(M) \approx \text{prior}(\hat{\theta}) \frac{1}{2\sqrt{2}} \frac{|X^T X|^{-1/2}}{\pi^{(n-p)/2}} \frac{1}{\|Y - X\hat{\beta}\|^{n-p-1}} \Gamma\left(\frac{n-p-1}{2}\right). \tag{2}$$

This formula is valid for any prior that is nearly constant (flat) in the region of concentration of the likelihood. Note that, when p is much smaller than n, the likelihood is quite concentrated around the MLE $\hat{\theta}$.

The Evidence depends on the prior through the value of the prior density at the MLE $\hat{\theta}$. In the last lecture, we worked with the prior:

$$\beta_0, \beta_1, \ldots, \beta_{2k}, \log \sigma \overset{\text{i.i.d}}{\sim} \text{Unif}(-C, C) \tag{3}$$

for a large C. This is not a good prior for linear regression because the β_j's can have quite different meanings and scales (e.g., X_1 might be age, X_2 might be current weight in pounds, X_3 might be weight a year ago in kilograms etc.) so using the same prior for all of them is not a good idea.

A better prior that is used in linear regression is that β and σ are independent with

$$\beta \sim N(0, \tau^2 (X'X)^{-1}) \quad \text{and} \quad \log \sigma \sim \text{Unif}(-C, C). \tag{4}$$

This prior depends on the two hyperparameters τ and σ. The normality assumption for β is standard and facilitates computation. The covariance of β is taken to be proportional to $(X'X)^{-1}$ as opposed to the identity matrix which means that the different coefficients β_j have different variances. The prior $\beta \sim N(0, \tau^2 (X'X)^{-1})$ is sometimes known as the Zellner prior (after the Econometrist Arnold Zellner).

2 Scaling Invariance of the Zellner Prior

The Zellner prior has the following very important scaling invariance property. Suppose we use transform our covariates and work with $\tilde{X} = XH$ for a nonsingular $p \times p$ matrix H. In other words, we are now working with the model:

$$Y = \tilde{X}\tilde{\beta} + Z \qquad \text{with } Z_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2) \tag{5}$$

(1) and (5) represent the same model because there is a one-to-one correspondence between β in (1) and $\tilde{\beta}$ in (5): $\tilde{\beta} = H^{-1}\beta$ (note then that $X\beta = \tilde{X}\tilde{\beta}$).

There are now two ways of choosing a prior for the parameter $\tilde{\beta}$ in the model (5). The first way is to use the Zellner prior in the model (5) which leads to:

$$\tilde{\beta} \sim N\left(0, \tau^2\left(\tilde{X}'\tilde{X}\right)^{-1}\right) = N\left(0, \tau^2\left(H'X'XH\right)^{-1}\right) = N\left(0, \tau^2 H^{-1}\left(X'X\right)^{-1}\left(H^{-1}\right)'\right)$$

The second way is to use the Zellner prior in model (1) for β and then use the equation $\tilde{\beta} = H^{-1}\beta$ to get the prior for $\tilde{\beta}$. This gives:

$$\beta \sim N\left(0, \tau^2(X'X)^{-1}\right) \quad \tilde{\beta} = H^{-1}\beta \implies \tilde{\beta} \sim N\left(0, \tau^2 H^{-1}(X'X)^{-1}(H^{-1})'\right).$$

Observe that both the above processes for choosing the prior for $\tilde{\beta}$ lead to exactly the same answer. This is the advantage of working with the Zellner prior.

3 Evidence for the prior (4)

We now calculate the Evidence of the linear regression model (1) for the prior (4). The prior density at $\theta = (\beta, \sigma)$ is given by

$$\begin{aligned}
\text{prior}(\theta) &= \text{prior}(\beta) \times \text{prior}(\sigma) \\
&= \left(\frac{1}{\sqrt{2\pi}\tau}\right)^p |X'X|^{1/2} \exp\left(-\frac{\beta' X'X\beta}{2\tau^2}\right) \frac{I\{-C < \log\sigma < C\}}{2C\sigma} \\
&= \left(\frac{1}{\sqrt{2\pi}\tau}\right)^p |X'X|^{1/2} \exp\left(-\frac{\|X\beta\|^2}{2\tau^2}\right) \frac{I\{e^{-C} < \sigma < e^C\}}{2C\sigma}.
\end{aligned}$$

Plugging this into the formula (2) then gives

$$\begin{aligned}
\text{Evidence}(M) &= \left(\frac{1}{\sqrt{2\pi}\tau}\right)^p |X'X|^{1/2} \exp\left(-\frac{\|X\hat{\beta}\|^2}{2\tau^2}\right) \frac{I\{e^{-C} < \hat{\sigma} < e^C\}}{2C\hat{\sigma}} \frac{1}{2\sqrt{2}} \frac{|X^TX|^{-1/2}}{\pi^{(n-p)/2}} \frac{\Gamma\left(\frac{n-p-1}{2}\right)}{\|Y - X\hat{\beta}\|^{n-p-1}} \\
&= \tau^{-p} \exp\left(-\frac{\|X\hat{\beta}\|^2}{2\tau^2}\right) \frac{I\{e^{-C} < \hat{\sigma} < e^C\}}{C\hat{\sigma}} \frac{2^{-(p+5)/2}}{\pi^{n/2}} \frac{\Gamma\left(\frac{n-p-1}{2}\right)}{\|Y - X\hat{\beta}\|^{n-p-1}}.
\end{aligned}$$

Here $\hat{\beta}$ and $\hat{\sigma}$ are the MLEs of β and σ respectively in the model M. Specifically, $\hat{\beta}$ is the least squares estimator:

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

and $\hat{\sigma}$ is given by

$$\hat{\sigma} = \frac{\|Y - X\hat{\beta}\|}{\sqrt{n}}$$

Plugging this value of $\hat{\sigma}$ in the Evidence formula, we obtain

$$\text{Evidence}(M) = \tau^{-p} \exp\left(-\frac{\|X\hat{\beta}\|^2}{2\tau^2}\right) \frac{I\{e^{-C} < \hat{\sigma} < e^C\}}{C} \frac{2^{-(p+5)/2}}{\pi^{n/2}} \frac{\Gamma\left(\frac{n-p-1}{2}\right)}{\|Y - X\hat{\beta}\|^{n-p}}.$$

This formula for the Evidence depends on the two quantities τ and C which appear in the specification of the prior (4). It is therefore prudent to use the notation Evidence$(M \mid \tau, C)$ to indicate this dependence explicitly:

$$\text{Evidence}(M \mid \tau, C) = \tau^{-p} \exp\left(-\frac{\|X\hat{\beta}\|^2}{2\tau^2}\right) \frac{I\{e^{-C} < \hat{\sigma} < e^C\}}{C} \frac{2^{-(p+5)/2}}{\pi^{n/2}} \frac{\Gamma\left(\frac{n-p-1}{2}\right)}{\|Y - X\hat{\beta}\|^{n-p}}.$$

The dependence of Evidence$(M \mid \tau, C)$ on C is quite mild because the indicator function is always one (as C will be large) and the C in the denominator will be the same for each linear model being considered. The dependence on τ, on the other hand, can be quite crucial as different values of τ might lead to quite different Evidences. To deal with this problem in a principled way, we integrate Evidence$(M \mid \tau, C)$ with respect to the following prior on τ:

$$\log \tau \sim \text{Unif}(-C_1, C_1)$$

for a large C_1. We then get

$$\begin{aligned}
\text{Evidence}(M \mid C) &= \int_0^\infty \text{Evidence}(M \mid \tau, C) f_\tau(\tau) d\tau \\
&= \int_{e^{-C_1}}^{e^{C_1}} \text{Evidence}(M \mid \tau, C) \frac{1}{2C_1 \tau} d\tau \\
&= \frac{1}{2C_1} \int_{e^{-C_1}}^{e^{C_1}} \text{Evidence}(M \mid \tau, C) \frac{1}{\tau} d\tau \\
&= \frac{1}{2C_1} \left[\int_{e^{-C_1}}^{e^{C_1}} \tau^{-p} \exp\left(-\frac{\|X\hat{\beta}\|^2}{2\tau^2}\right) \frac{d\tau}{\tau}\right] \frac{I\{e^{-C} < \hat{\sigma} < e^C\}}{C} \frac{2^{-(p+5)/2}}{\pi^{n/2}} \frac{\Gamma\left(\frac{n-p-1}{2}\right)}{\|Y - X\hat{\beta}\|^{n-p}} \\
&\approx \frac{1}{2C_1} \left[\int_0^\infty \tau^{-p-1} \exp\left(-\frac{\|X\hat{\beta}\|^2}{2\tau^2}\right) d\tau\right] \frac{I\{e^{-C} < \hat{\sigma} < e^C\}}{C} \frac{2^{-(p+5)/2}}{\pi^{n/2}} \frac{\Gamma\left(\frac{n-p-1}{2}\right)}{\|Y - X\hat{\beta}\|^{n-p}}
\end{aligned}$$

The integral above can be evaluated in closed form using the substitution:

$$s = \frac{\|X\hat{\beta}\|^2}{2\tau^2} \quad \text{or} \quad \tau = \frac{\|X\hat{\beta}\|}{\sqrt{2s}} \implies d\tau = -\frac{\|X\hat{\beta}\|}{2\sqrt{2}} s^{-3/2} ds$$

which leads to

$$\int_0^\infty \tau^{-p-1} \exp\left(-\frac{\|X\hat{\beta}\|^2}{2\tau^2}\right) d\tau = \frac{2^{(p-2)/2}}{\|X\hat{\beta}\|^p} \int_0^\infty s^{\frac{p}{2}-1} e^{-s} ds = 2^{(p-2)/2} \frac{\Gamma(p/2)}{\|X\hat{\beta}\|^p}.$$

We have therefore obtained

$$\begin{aligned}
\text{Evidence}(M \mid C) &= \frac{1}{2C_1} 2^{(p-2)/2} \frac{\Gamma(p/2)}{\|X\hat{\beta}\|^p} \frac{I\{e^{-C} < \hat{\sigma} < e^C\}}{C} \frac{2^{-(p+5)/2}}{\pi^{n/2}} \frac{\Gamma\left(\frac{n-p-1}{2}\right)}{\|Y - X\hat{\beta}\|^{n-p}} \\
&= \frac{1}{2^{9/2}} \frac{I\{e^{-C} < \hat{\sigma} < e^C\}}{CC_1 \pi^{n/2}} \frac{\Gamma(p/2)}{\|X\hat{\beta}\|^p} \frac{\Gamma\left(\frac{n-p-1}{2}\right)}{\|Y - X\hat{\beta}\|^{n-p}}.
\end{aligned}$$

This Evidence depends on C as well as on C_1 so it will be better to call this Evidence($M \mid C, C_1$):

$$\text{Evidence}(M \mid C, C_1) = \frac{1}{2^{9/2}} \frac{I\{e^{-C} < \hat{\sigma} < e^{C}\}}{CC_1\pi^{n/2}} \frac{\Gamma(p/2)}{\|X\hat{\beta}\|^p} \frac{\Gamma\left(\frac{n-p-1}{2}\right)}{\|Y - X\hat{\beta}\|^{n-p}}.$$

When we are performing model selection among different linear models and use the same C and C_1 for all of them, then the precise choices of C and C_1 will have no effect as they will only modify each Evidence by a constant amount. In fact, we can ignore terms which do not depend on the specific model M and write:

$$\text{Evidence}(M) \propto \frac{\Gamma\left(\frac{p}{2}\right)}{\|X\hat{\beta}\|^p} \frac{\Gamma\left(\frac{n-p-1}{2}\right)}{\|Y - X\hat{\beta}\|^{n-p}}. \tag{6}$$

The above expression does not depend on any tuning parameters. We can calculate this for each individual linear model and then normalize the evidences so they sum to one (this lets us interpret them as posterior probabilities of the different models). In the rest of this lecture, we shall illustrate the performance of this method on various real and simulated datasets.

4 Recommended List of Readings for Today

1. The Evidence formula (6) essentially comes from Chapter 5 of the book *Bayesian spectrum analysis and parameter estimation* by Larry Bretthorst, and
2. Another way to calculate the Evidence for linear regression models is described in Section 9.3.1 of the book *A first course in Bayesian Statistical Methods* by Peter Hoff. Here they use a prior that is slightly different from (4). Specifically, they assume that $\tau^2 = g\sigma^2$ (which makes β and σ dependent).