STAT 153 & 248 - Time Series
Lecture Six

Fall 2025, UC Berkeley

Aditya Guntuboyina

September 16, 2025

1 Nonlinear Regression

We started discussing nonlinear regression models near the end of last lecture. In these models, certain parameters appear in a nonlinear fashion. One simple example of such a model is:

$$y_t = \beta_0 + \beta_1 t + \beta_2 \text{ReLU}(t - c) + \epsilon_t \tag{1}$$

with $\epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. Here $\text{ReLU}(t-c) = (t-c)_+$ equals 0 if $t \leq c$ and equals $t - c$ if $t \geq c$. We can also write

$$\text{ReLU}(t - c) = (t - c)_+ = (t - c)I\{t > c\} = \max(t - c, 0).$$

$(\cdot)_+$ is also called the positive part function, or, the ramp function.

The model (1) says that for times $t \leq c$, the slope of the regression line is β_1, while for $t > c$, the slope changes to $(\beta_1 + \beta_2)$. We shall refer to (1) as the 'Change of Slope' model. An alternative name for this model is "Broken-stick regression". This is because the function

$$t \mapsto \beta_0 + \beta_1 t + \beta_2 \text{ReLU}(t - c)$$

resembles a broken stick.

The unknown parameters for this model are $c, \beta_0, \beta_1, \beta_2$ as well as σ. The unknown parameter c makes (1) a nonlinear regression model. If c were known, then (1) would be a linear regression model:

$$y = X_c \beta + \epsilon \tag{2}$$

with

$$X_c = \begin{pmatrix} 1 & 1 & \text{ReLU}(1-c) \\ 1 & 2 & \text{ReLU}(2-c) \\ 1 & 3 & \text{ReLU}(3-c) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & n & \text{ReLU}(n-c) \end{pmatrix} \text{ and } \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \text{ and } \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{pmatrix}$$

1.1 Parameter Estimation

Least squares again is the most basic estimation procedure. The sum of squares is:

$$S(\beta_0, \beta_1, \beta_2, c) := \sum_{t=1}^{n} \left(y_t - \beta_0 - \beta_1 t - \beta_2 \text{ReLU}(t-c)\right)^2. \tag{3}$$

We need to minimize this over all the four variables $\beta_0, \beta_1, \beta_2, c$. Using matrix notation, we can write

$$S(\beta, c) = \|y - X_c \beta\|^2.$$

If we fix c, then it is easy to minimize $S(\beta, c)$ over β. This is the same as linear regression and the minimizing β is given by:

$$\hat{\beta}(c) := (X_c^T X_c)^{-1} X_c^T y, \tag{4}$$

and the smallest value of $S(\beta, c)$ for fixed c is $S(\hat{\beta}(c), c)$ which is just the RSS in the multiple linear regression with fixed c. We use the notation:

$$RSS(c) = S(\hat{\beta}(c), c) = \min_{\beta} S(\beta, c). \tag{5}$$

The least squares estimates of β and c can therefore be found in the following way:

1. Fix a finite set of possible values of c. In this change of slope model, it is reasonable to assume that $c \in \{1, \ldots, n\}$. One can also take a finer grid of values in the range $[1, n]$.
2. For each value of c in the chosen set, calculate $\hat{\beta}(c)$ and define $RSS(c) = S(\hat{\beta}(c), c)$.
3. Take $\hat{c}$ to be the value of c which minimizes $RSS(c)$.
4. Take $\beta = \hat{\beta}(\hat{c})$.

1.2 Bayesian Inference

For uncertainty quantification, let us consider Bayesian inference for this model. We need to first write down the likelihood and prior, and then use them to derive the posterior. For the likelihoood, we shall assume (just as in usual linear regression) that the normal distribution for the errors $\{\epsilon_t\}$ in (1): $\epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. This leads to the likelihood:

$$\prod_{t=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_t - \beta_0 - \beta_1 t - \beta_t (t-c)_+)^2}{2\sigma^2}\right) = (2\pi)^{-n/2} \sigma^{-n} \exp\left(-\frac{S(\beta, c)}{2\sigma^2}\right) \tag{6}$$

where again $S(\beta, c)$ is the sum of squares (3).

For the prior, recall that, for linear regression, we used:

$$\beta_0, \beta_1, \beta_2, \log \sigma \overset{\text{i.i.d}}{\sim} \text{unif}(-\infty, \infty).$$

Here $\text{unif}(-\infty, \infty)$ can be thought of as short form for $\text{unif}(-C, C)$ for a very large constant C. We will use the same prior here also. However we also need to supply a prior for the parameter c. Here note that c needs to be in the range $(1, n)$ (open interval from 1 to n). Because if $c \geq n$, then $\text{ReLU}(t-c) = 0$ for all $t = 1, \ldots, n$ so the term $\text{ReLU}(t-c)$ is not needed in the model (1). If $c \leq 1$, then $\text{ReLU}(t-c) = t - c$ for all $t = 1, \ldots, n$ so that the ReLU term in (1) becomes linear so the model again becomes a linear function of t.

Because $c \in (1, n)$, a natural prior is to take:

$$c \sim \text{uniform}(1, n). \tag{7}$$

In practice, we would not be interested in values of c that are near the boundaries 1 and n. This is because if c is very close to 1 or n, then the model is essentially linear for most of dataset. To reflect this, we can take the prior for c to be uniform in smaller range than $(1, n)$ ignoring points near each end; for example, uniform on $(5, n-5)$. Let us work with the prior (7), and we shall mention the changes that need to be made to the posterior if we are instead working with the uniform prior on a smaller range such as $(5, n-5)$.

The posterior joint density of all the parameters β, c, σ (here β denotes the vector consisting of $\beta_0, \beta_1\beta_2$) is given by

$$\text{posterior}(\beta, c, \sigma) \propto \text{likelihood} \times \text{prior}.$$

The likelihood is given in (6) and the prior is

$$\begin{aligned}\text{prior density} &= \frac{I\{-C < \beta_0, \beta_1, \beta_2, \log\sigma < C\}}{(2C)^4\sigma} \frac{I\{1 < c < n\}}{n-1} \\ &\propto \frac{I\{-C < \beta_0, \beta_1, \beta_2, \log\sigma < C, 1 < c < n\}}{\sigma}\end{aligned}$$

We shall drop the indicator terms involving C because C is very large (think ∞). This will lead to

$$\text{prior density} \propto \frac{I\{\sigma > 0, 1 < c < n\}}{\sigma}$$

The posterior is then given by

$$\text{posterior}(\beta, c, \sigma) \propto \sigma^{-n-1} \exp\left(-\frac{S(\beta, c)}{2\sigma^2}\right) I\{\sigma > 0\} I\{1 < c < n\}.$$

To get the posterior density of c alone (c is the most important parameter in the model (1)), we need to integrate the joint posterior density above with respect to β and σ:

$$\text{posterior}(c) \propto I\{1 < c < n\} \int_0^\infty \sigma^{-n-1} \int \exp\left(-\frac{S(\beta, c)}{2\sigma^2}\right) d\beta d\sigma.$$

Let us first calculate the inner integral. $S(\beta, c)$ is a quadratic function in β so that $\int \exp(-S(\beta, c)/(2\sigma^2))d\beta$ should be related to the normalizing constants in the multivariate normal density. To figure the integral precisely, we first use the Pythagorean identity (discussed previously):

$$\begin{aligned}S(\beta, c) &= S(\hat\beta(c), c) + \left(\beta - \hat\beta(c)\right)^T X_c^T X_c \left(\beta - \hat\beta(c)\right) \\ &= RSS(c) + \left(\beta - \hat\beta(c)\right)^T X_c^T X_c \left(\beta - \hat\beta(c)\right)\end{aligned}$$

where $\hat\beta_c$ is given in (4). The value $S(\hat\beta_c, c)$ is equal to $RSS(c)$ as noted in (5). Thus

$$\begin{aligned}\int \exp\left(-\frac{S(\beta, c)}{2\sigma^2}\right) d\beta &= \int \exp\left(-\frac{RSS(c)}{2\sigma^2}\right) \exp\left(-\frac{\left(\beta - \hat\beta(c)\right)^T X_c^T X_c \left(\beta - \hat\beta(c)\right)}{2\sigma^2}\right) d\beta \\ &= \exp\left(-\frac{RSS(c)}{2\sigma^2}\right) \int \exp\left(-\frac{\left(\beta - \hat\beta(c)\right)^T X_c^T X_c \left(\beta - \hat\beta(c)\right)}{2\sigma^2}\right) d\beta \\ &= \exp\left(-\frac{RSS(c)}{2\sigma^2}\right) (\sqrt{2\pi})^p \sqrt{\det\left(\sigma^2 (X_c^T X_c)^{-1}\right)} \\ &= \exp\left(-\frac{RSS(c)}{2\sigma^2}\right) (\sqrt{2\pi})^p \sigma^p \left|X_c^T X_c\right|^{-1/2}\end{aligned}$$

where $|X_c^T X_c| = \det(X_c^T X_c)$. Here $p = 3$ because there are three components inside β. Therefore

$$\begin{aligned}\text{posterior}(c) &\propto I\{1 < c < n\} \int_0^\infty \sigma^{-n-1} \exp\left(-\frac{RSS(c)}{2\sigma^2}\right) (\sqrt{2\pi})^p \sigma^p \left|X_c^T X_c\right|^{-1/2} d\sigma \\ &\propto I\{1 < c < n\} \left|X_c^T X_c\right|^{-1/2} \int_0^\infty \sigma^{-n+p-1} \exp\left(-\frac{RSS(c)}{2\sigma^2}\right) d\sigma.\end{aligned}$$

The change of variable $\sigma = s\sqrt{RSS(c)}$, gives

$$\begin{aligned}\text{posterior}(c) &\propto I\{1 < c < n\} \left|X_c^T X_c\right|^{-1/2} \left(\frac{1}{RSS(c)}\right)^{(n-p)/2} \int_0^\infty s^{-n+p-1} \exp\left(-\frac{1}{2s^2}\right) dt \\ &\propto I\{1 < c < n\} \left|X_c^T X_c\right|^{-1/2} \left(\frac{1}{RSS(c)}\right)^{(n-p)/2}. \qquad (8)\end{aligned}$$

This posterior will be evaluated numerically over a grid of values of c in the range $(1, n)$. The term $|X_c^T X_c|^{-1/2}$ will become infinite when $|X_c^T X_c| = 0$ i.e., when X_c does not have full column rank. This is the case when c is outside of the range $(1, n)$. When c is in the range $(1, n)$, the determinant of $X_c^T X_c$ will be non-zero but it will still be good to not consider c too close to 1 or n to prevent numerical instability due to near-singularity.

If the prior is taken to be uniform on a different range (say $(5, n-5)$), then we simply need to change the indicator $I\{1 < c < n\}$ in (8) by the indicator over the prior range.

The main term in the posterior (8) is $(1/RSS(c))^{(n-p)/2}$ (the other term $|X_c^T X_c|^{-1/2}$ generally does not vary significantly with c). This term takes its largest value when c equals the least squares estimator $\hat{c}$ (note $\hat{c}$ minimizes $RSS(c)$). The size of the power $n - p$ determines the amount of concentration of the posterior around $\hat{c}$. When $n - p$ is large, the posterior is very tightly concentrated around $\hat{c}$.