STAT 153 & 248 - Time Series Lecture Nine

Spring 2025, UC Berkeley

Aditya Guntuboyina

February 18, 2025

1 Change of Slope Model

While our focus, in the last few lectures, has been on sinusoidal models, the methodology can be applied in the same way to some other nonlinear regression models. As an illustrative example, we study the change of slope model today. Other examples can be found in Homework Two. The change of slope model is given by:

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

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 $(\beta_1 + \beta_2)$. 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

$$X_c = \begin{pmatrix} 1 & 1 & \text{ReLU}(1-c) \\ 1 & 2 & \text{ReLU}(2-c) \\ 1 & 3 & \text{ReLU}(3-c) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & n & \text{ReLU}(n-c) \end{pmatrix}$$

2 Estimation of $c, \beta_0, \beta_1, \beta_2, \sigma$

Parameter estimation can be done via Maximum Likelihood Estimation. Just as in the case of the sinusoidal model, a crucial role is played by the residual sum of squares in the linear regression model (2):

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

The MLE of c is given by the minimizer of $RSS(c)$ over c. On the computer, we calculate this by enumerating all the possible values of c (these are just $1, 2, \dots, n$) and then using a function such as `np.argmin`.

After the MLE $\hat{c}$ of c is obtained, $\beta_0, \beta_1, \beta_2$ are estimated as in linear regression (e.g., using `sm.OLS`). Specifically,

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)^T = (X_{\hat{c}}^T X_{\hat{c}})^{-1} X_{\hat{c}}^T y. \tag{3}$$

Then σ is estimated as in linear regression:

$$\hat{\sigma} = \sqrt{\frac{RSS(\hat{c})}{n-3}}.$$

3 Uncertainty Quantification for $c, \beta_0, \beta_1, \beta_2, \sigma$

For this, we use Bayesian analysis. Our prior for $\beta_0, \beta_1, \beta_2, \sigma$ is the same as the one used for linear regression:

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

for a large C.

For the parameter c, we also use a uniform prior. The range of values of c is $1, 2, \dots, n$. But actually there is no reason to allow $c = 1$ and $c = n$ as explained below.

When $c = 1$, the variable $\text{ReLU}(t-c)$ simply becomes $t - c = t - 1$ so that the nonlinear term $\text{ReLU}(t-c)$ can be absorbed with the other terms as:

$$\beta_0 + \beta_1 t + \beta_2 \text{ReLU}(t-1) = \beta_0 + \beta_1 t + \beta_2 (t-1) = (\beta_0 - \beta_2) + (\beta_1 + \beta_2) t.$$

In other words, when $c = 1$, the model reverts to the simple linear trend model (it is no longer a broken stick regression model). On the other hand, when $c = n$, we simply have $\text{ReLU}(t-c) = 0$ (because $t = 1, \dots, n$ is always smaller than n) so the term $\text{ReLU}(t-c)$ has no effect when $c = n$.

Our range of values is therefore $c = 2, \dots, n$. The prior for c will be taken to be

$$c \sim \text{uniform}\{2, \dots, n-1\}.$$

With these priors, calculate the posterior distributions, exactly as in the case of the sinusoidal model. This leads to the following. The posterior distribution of c is:

$$\pi(c \mid \text{data}) \propto \left(\frac{1}{RSS(c)}\right)^{(n-3)/2} |X_c^T X_c|^{-1/2} I\{c = 2, \dots, n-1\}.$$

In other words, this is a discrete distribution with pmf:

$$\pi(c \mid \text{data}) = \frac{\left(\frac{1}{RSS(c)}\right)^{(n-3)/2} |X_c^T X_c|^{-1/2}}{\sum_{c=2}^{n-1} \left(\frac{1}{RSS(c)}\right)^{(n-3)/2} |X_c^T X_c|^{-1/2}} \qquad \text{for } c = 2, \dots, n-1.$$

The denominator above is simply the sum of the numerator values for all $c = 2, \dots, n-1$. In particular, the denominator does not depend on the particular value of c anymore and is a constant. We can also write:

$$\pi(c \mid \text{data}) = \frac{\left(\frac{1}{RSS(c)}\right)^{(n-3)/2} |X_c^T X_c|^{-1/2}}{\left(\frac{1}{RSS(2)}\right)^{(n-3)/2} |X_2^T X_2|^{-1/2} + \dots + \left(\frac{1}{RSS(n-1)}\right)^{(n-3)/2} |X_{n-1}^T X_{n-1}|^{-1/2}}$$

for $c = 2, \dots, n-1$.

Given c, as remarked before, the model is just a linear regression model with X-matrix given by X_c. Therefore, by results from linear regression (see Problem 4 in Homework 1), the posterior density of σ given the data as well as c is characterized by:

$$\frac{RSS(c)}{\sigma^2} \mid \text{data}, c \sim \chi^2_{n-3}.$$

Finally, the posterior distribution of β given the data as well as c and σ is

$$\beta \mid \text{data}, c, \sigma \sim N_3\left(\hat{\beta}_c, \sigma^2 (X_c^T X_c)^{-1}\right),$$

where

$$\hat{\beta}_c := (X_c^T X_c)^{-1} X_c^T y.$$

4 Posterior Sampling for Uncertainty Quantification

A useful way of visualizing the uncertainty is to draw posterior samples from the unknown parameters, and then plot the corresponding fitted values along with the observed data. The algorithm for drawing the posterior samples is as follows.

1. Obtain samples $c^{(1)}, \dots, c^{(N)}$ by sampling, with replacement, from the set of possible values of $c :\ \{2, \dots, n\}$ with probability weights given by the posterior pmf $\pi(c \mid \text{data})$. For this, one can use, for example, the choice function in `np.random.default_rng()`.

2. For each $j = 1, \dots, N$,

   a) Fix $c = c^{(j)}$.

   b) Calculate $RSS(c)$ and $\hat{\beta}_c$ by implementing linear regression with fixed c.

   c) Generate a chi-squared random variable χ^2 with $n-3$ degrees of freedom. Take $\sigma^{(j)} = \sqrt{RSS(c)/\chi^2}$.

   d) Take $\beta^{(j)}$ to be a generated random vector from the multivariate normal distribution with mean $\hat{\beta}_c$ and covariance $(\sigma^{(j)})^2 (X_c^T X_c)^{-1}$.

Suppose the posterior samples are given by $(c^{(j)}, \beta_0^{(j)}, \beta_1^{(j)}, \beta_2^{(j)}, \sigma^{(j)})$ for $j = 1, \dots, N$. The corresponding fitted values are given by:

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

for $t = 1, \dots, n$. These can be plotted along with the original data.

One can also plot vertical lines corresponding to $c^{(j)}$ to visualize uncertainty in the change-of-slope time point parameter c.

The 95% approximate credible intervals for each parameter can be obtained by computing the 2.5th and 97.5th percentiles of the corresponding posterior samples.

Suppose you want to obtain posterior samples for y_{t^*} for a future time point t^*. One can follow the algorithm listed above with one additional step to generate $y_{t^*}^{(j)}, j = 1, \dots, N$ as follows:

1. Obtain samples $c^{(1)}, \dots, c^{(N)}$ by sampling, with replacement, from the set of possible values of $c : \{2, \dots, n\}$ with probability weights given by the posterior pmf $\pi(c \mid \text{data})$. For this, one can use, for example, the choice function in `np.random.default_rng()`.
2. For each $j = 1, \dots, N$,
   a) Fix $c = c^{(j)}$.
   b) Calculate $RSS(c)$ and $\hat{\beta}_c$ by implementing linear regression with fixed c.
   c) Generate a chi-squared random variable χ^2 with $n - 3$ degrees of freedom. Take $\sigma^{(j)} = \sqrt{RSS(c)/\chi^2}$.
   d) Take $\beta^{(j)}$ to be a generated random vector from the multivariate normal distribution with mean $\hat{\beta}_c$ and covariance $(\sigma^{(j)})^2(X_c^T X_c)^{-1}$.
   e) Generate $y_{t^*}^{(j)}$ from the normal distribution with mean $\beta_0^{(j)} + \beta_1^{(j)} t^* + \beta_2^{(j)} \text{ReLU}(t^* - c^{(j)})$ and variance $(\sigma^{(j)})^2$.

5 More Changes of Slope

Suppose we want to introduce two break points for the regression line. This can be done via:

$$y_t = \beta_0 + \beta_1 t + \beta_2 \text{ReLU}(t - c_1) + \beta_3 \text{ReLU}(t - c_2) + \epsilon_t$$

with $\epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. This model can also be written as

$$y = X_c \beta + \epsilon$$

where $c = (c_1, c_2)$ and

$$X_c := \begin{pmatrix} 1 & 1 & \text{ReLU}(1 - c_1) & \text{ReLU}(1 - c_2) \\ 1 & 2 & \text{ReLU}(2 - c_1) & \text{ReLU}(2 - c_2) \\ 1 & 3 & \text{ReLU}(3 - c_1) & \text{ReLU}(3 - c_2) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & n & \text{ReLU}(n - c_1) & \text{ReLU}(n - c_2) \end{pmatrix}$$

The MLE of $c = (c_1, c_2)$ is given by the minimizer of $RSS(c)$ over all $c = (c_1, c_2)$ with $c_1, c_2 \in \{1, \dots, n\}$, and

$$RSS(c) = \min_{\beta} \|y - X_c \beta\|^2.$$

One can numerically minimize $RSS(c)$ over all $c_1, c_2 \in \{1, \ldots, n\}$. The posterior of c becomes:

$$\pi(c \mid \text{data}) \propto \left(\frac{1}{RSS(c)}\right)^{(n-4)/2} |X_c^T X_c|^{-1/2}.$$

For more break points, one can consider:

$$y_t = \beta_0 + \beta_1 t + \sum_{j=1}^{k} \beta_{j+1} \text{ReLU}(t - c_j) + \epsilon_t.$$

Conceptually estimation and inference here proceed just as before with X_c changed appropriately. But the method can become computationally expensive if $k \geq 4$.