STAT 153 & 248 - Time Series
Lecture Ten

Spring 2025, UC Berkeley

Aditya Guntuboyina

February 20, 2025

For a given time series $y_1, \ldots, y_n$, we consider the model:

$$y_t = \beta_0 + \beta_1(t-1) + \beta_2 \text{ReLU}(t-2) + \cdots + \beta_{n-1}\text{ReLU}(t-(n-1)) + \epsilon_t \quad (1)$$

where, as always, $\epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. Here $\text{ReLU}(t-c) = (t-c)_+$ equals 0 if $t \leq c$ and equals $(t-c)$ if $t > c$.

The unknown parameters in this model are $\beta_0, \beta_1, \ldots, \beta_{n-1}$ as well as σ. The model (1) should be compared with the following model that we studied in the previous lecture:

$$y_t = \beta_0 + \beta_1(t-1) + \beta_2 \text{ReLU}(t-c_1) + \beta_3 \text{ReLU}(t-c_2) + \cdots + \beta_{k+1}\text{ReLU}(t-c_k) + \epsilon_t. \quad (2)$$

Here are the main differences between these two models:

1. The model (2) will be used with a small value of k (such as $1, 2, 3, 4$). This makes it a low-dimensional model. On the other hand, the number of unknown parameters in (1) equals $n+1$ which is quite large. So (1) is an example of a high-dimensional model.
2. (2) is a nonlinear model because of the presence of the parameters $c_1, \ldots, c_k$. On the other hand, there are no such nonlinear parameters in (1) which makes it a linear regression model.

To summarize, (2) is a low-dimensional nonlinear regression model, while (1) is a high-dimensional linear regression model.

1 Parameter Interpretation in (1)

Let μ_t denote the deterministic part of model (1), i.e.,

$$\mu_t = \beta_0 + \beta_1(t-1) + \beta_2 \text{ReLU}(t-2) + \beta_3 \text{ReLU}(t-3) + \cdots + \beta_{n-1}\text{ReLU}(t-(n-1)). \quad (3)$$

The model (1) can then be written as

$$y_t = \mu_t + \epsilon_t \qquad \text{with } \epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2).$$

Here μ_t represents the trend function that we aim to estimate from the data. For illustration, consider y_t to be the logarithm of California's population in year t. The trend function μ_t captures the underlying systematic pattern in the population growth, while ϵ_t accounts for random fluctuations around this trend.

Sometimes, we can also interpret μ_t as the 'actual' data and ϵ_t as the measurement error causing μ_t to be observed as y_t. For example, in the population example, μ_t would represent the actual population (on log scale) while y_t would represent our noisy measurement of it.

The parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_{n-1}$ can be interpreted in terms of μ_t as follows. We will focus on the population example here for simplicity. Plugging $t = 1$ in (3), we get

$$\beta_0 = \mu_1. \tag{4}$$

So β_0 can be interpreted as the actual population on log scale (or the value of the trend function) at time $t = 1$. Plugging $t = 2$ in (3), we get $\mu_2 = \beta_0 + \beta_1 = \mu_1 + \beta_1$ so that

$$\beta_1 = \mu_2 - \mu_1.$$

If $P_t = \exp(\mu_t)$ denotes the population on the original scale, then

$$\beta_1 = \log P_2 - \log P_1 = \log \frac{P_2}{P_1} \approx \frac{P_2}{P_1} - 1 = \frac{P_2 - P_1}{P_1}.$$

Here we used the fact that $\log x \approx x - 1$ if x is close to 1. In other words, $100\beta_1$ can be interpreted as the percentage growth of the population from year 1 to year 2.

For β_2, let us plug $t = 3$ in (3) to get $\mu_3 = \beta_0 + 2\beta_1 + \beta_2$. Replacing $\beta_0 = \mu_1$ and $\beta_1 = \mu_2 - \mu_1$, we obtain

$$\beta_2 = (\mu_3 - \mu_2) - (\mu_2 - \mu_1).$$

This means that

$100\beta_2 \approx$ (percentage change from year 2 to 3) $-$ (percentage change from year 1 to 2)

Continuing this way for $t = 4, 5, ..., n$, we get

$$\beta_t = (\mu_{t+1} - \mu_t) - (\mu_t - \mu_{t-1})$$

so that

$100\beta_t \approx$ (percentage change from year t to $(t+1)$)$-$(percentage change from year $(t-1)$ to t).

For example, suppose

$$\beta_0 = 7.3 \quad \beta_1 = 0.04 \quad \beta_2 = -0.001 \quad \beta_3 = -0.0005 \text{ etc.}$$

The interpretation then is that μ_t started with the value $\mu_1 = \exp(7.3) \approx 1480$ (if the population units are in thousands of persons, this means that the population at time 1 was 1.48 million). From year 1 to year 2, the population grew by 4%. From year 2 to year 3, the population grew by $4 - 0.1 = 3.9\%$. From year 3 to year 4, the population grew by $3.9 - 0.05 = 3.85\%$, and so on.

It is important to understand that the parameters $\beta_2, \ldots, \beta_{n-1}$ are on a different scale (units) compared to β_0 and β_1. β_0 is in the scale of the data, $100\beta_1$ represents percent change, while $100\beta_j$ for $j \geq 2$ represents the change in percent change.

We next study strategies for estimating the unknown parameters $\beta_0, \beta_1, \ldots, \beta_{n-1}$ (as well as σ) from the observed time series $y_1 \ldots . y_n$.

2 (Unregularized) MLE

Since (1) is a linear regression model, we can estimate the coefficients in the usual way by the MLE, or equivalently, least squares by minimizing

$$\sum_{t=1}^{n} \left(y_t - \beta_0 - \beta_1(t-1) - \beta_2 \text{ReLU}(t-2) - \cdots - \beta_{n-1}\text{ReLU}(t-(n-1))\right)^2$$

over all $\beta_0, \ldots, \beta_{n-1}$. The smallest value achievable in the above minimization will be the RSS. The MLE of σ is then given by

$$\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{RSS}{n}}.$$

Since there are as many coefficients as there are data points, this approach will give a perfect fit to the data leading to $RSS = 0$. In fact, from the work done in the previous section, the values of $\beta_0, \ldots, \beta_{n-1}$ which minimize the sum of squares are given by:

$$\beta_0 = y_1 \quad \beta_1 = y_2 - y_1 \quad \beta_j = (y_{j+1} - y_j) - (y_j - y_{j-1})$$

for $j = 2, \ldots, n-1$. This will lead to the estimated trend function $\mu_t = y_t$ for all t. Also the MLE of σ will be zero. The unbiased estimate of σ (that we previousy used in linear regression) will not exist because it will equal $\sqrt{RSS/(n-p)}$ with $p = n$.

To summarize, these estimates will overfit the data, and will not produce a trend estimate that is simpler than the observed data.

3 Regularization

To produce useful estimates in cases where the MLE overfits, one employs the idea of regularization. We will discuss two ways of doing this: Ridge regularization and LASSO regularization.

The Ridge estimate of β will be denoted by $\hat{\beta}_{\text{ridge}}(\lambda)$ and is given by the minimizer of:

$$\begin{aligned}&\sum_{t=1}^{n} \left(y_t - \beta_0 - \beta_1(t-1) - \beta_2 \text{ReLU}(t-2) - \cdots - \beta_{n-1}\text{ReLU}(t-(n-1))\right)^2 \\ &+ \lambda\left(\beta_2^2 + \beta_3^2 + \cdots + \beta_{n-1}^2\right).\end{aligned} \tag{5}$$

In other words, $\hat{\beta}_{\text{ridge}}(\lambda)$ minimizes a new criterion function that is obtained by adding the penalty term $\lambda(\sum_{j=2}^{n-1} \beta_j^2)$ to the least squares criterion.

Here λ denotes a tuning parameter. Different choices of λ give rise to different ridge estimators $\hat{\beta}_{\text{ridge}}(\lambda)$. When $\lambda = 0$, the penalty term is not used in (5) so that $\hat{\beta}_{\text{ridge}}(\lambda)$ coincides with the unregularized least squares estimator. If λ is set to be very large, then the penalty term dominates the objective function (5) and then the first two components of $\hat{\beta}_{\text{ridge}}(\lambda)$ coincide with linear regression while the last $n-2$ components are simply set to zero.

The LASSO estimate of β will be denoted by $\hat{\beta}_{\text{lasso}}(\lambda)$ and is given by the minimizer of:

$$\begin{aligned}&\sum_{t=1}^{n} \left(y_t - \beta_0 - \beta_1(t-1) - \beta_2 \text{ReLU}(t-2) - \cdots - \beta_{n-1}\text{ReLU}(t-(n-1))\right)^2 \\ &+ \lambda\left(|\beta_2| + |\beta_3| + \cdots + |\beta_{n-1}|\right).\end{aligned} \tag{6}$$

In other words, $\hat{\beta}_{\text{lasso}}(\lambda)$ minimizes a new criterion function that is obtained by adding the penalty term $\lambda(\sum_{j=2}^{n-1} |\beta_j|)$ to the least squares criterion. As in the case of the ridge estimator, when $\lambda = 0$, the penalty term is not used in (6) so that $\hat{\beta}_{\text{ridge}}(\lambda)$ coincides with the unregularized least squares estimator. If λ is set to be very large, then the penalty term dominates the objective function (6) and then the first two components of $\hat{\beta}_{\text{ridge}}(\lambda)$ coincide with linear regression while the last $n-2$ components are simply set to zero.

The only difference between the ridge and lasso is in the penalty term: $\sum_j \beta_j^2$ vs $\sum_j |\beta_j|$. We will discuss computation and the differences between these estimators in the next lecture.

Note that, in usual implementations of ridge and lasso, the penalty is usually placed on all the coefficients (with the possible exception of the intercept). Here we are only placing it on $\beta_2, \ldots, \beta_{n-1}$. As we saw in the interpretation section, β_1 is quite different (both in having different units and also being somewhat bigger in size) compared to $\beta_2, \ldots, \beta_{n-1}$. It would not make sense in this example to include β_1 in the penalty term.