HDCC 预测模型（2）：Prophet 模型

每个模型其实都有自己的思想；或者换个说法，对每个模型的提出者来讲，模型背后都有他自己的逻辑表述。时间序列建模的思想是，在没有其他变量可用的时候，为了发现目标变量（Y）的规律，充分利用数据的生存过程（Data generate process, DGP），来构建 Y 的序列模型，进而对 Y 的未来发展趋势进行预测。

时间序列建模时，虽然没有其他变量（X）可用，但我们还可以考虑哪些方面的因素，进而有助于建模呢？归纳起来，主要有四个方面的因素：第一，Y 自身的发展规律，从而引出常见 AR 模型。第二，随机扰动项因素，从而引出常见的 MA 模型以及 ARCH 类模型。第三，时间(t)因素，从而引出 Y 对 t 的建模。第四，将 Y 分解为四大因素（T/C/S/I）,从季节调整的角度进行建模（当然，也有分解为五大因素一说，第五大因素即日历因素 D，具体可参考石刚（2013））。

常见的时间序列模型有 AR 模型，MA 模型，ARMA 模型，ARIMA 模型，SARIMA 模型，ARCH 类模型。多变量时间序列模型有 VAR 模型，SVAR 模型。季节调整类的模型主要有 X-13 模型（最新版到 X-13-ARIMA-SEAT），TRAMO/SEATS 模型，Prophet 模型。今天我们主要来了解一下 Prophet 模型。

Prophet 模型的基本思想是将序列 Y 分解为趋势（为了后面公式方便，这里用 g 表示），季节（s），节假日（h）以及随机项 e，然后对 g, s, h 这三项分别建模。

Prophet 趋势项中有两个重要的函数，分别为逻辑回归函数和分段线性函数，具体表述如下：

分段逻辑回归增长模型为：

$$g(t)=\frac{C(t)}{1+exp\left(-(k+\boldsymbol{a}(t)^T\boldsymbol{\delta})\right)(t-m+\boldsymbol{a}(t)^T\boldsymbol{\gamma})}$$

$$a_j(t)=\begin{cases}1, & t>s_j\\ 0, & otherwise\end{cases}$$

其中，$\boldsymbol{a}(t)=(a_1(t),\ldots,a_s(t))^T,\boldsymbol{\delta}=(\delta_1,\ldots,\delta_s)^T,\boldsymbol{\gamma}=(\gamma_1,\ldots,\gamma_s)^T$，$C(t)$为饱和值，在模型中需要提前设定其取值，$k+\boldsymbol{a}(t)^T\delta$为随时间变化的增长率，$m+\boldsymbol{a}(t)^T\boldsymbol{\gamma}$为偏置参数，$s_j$为序列中的突变点。

对于分段线性函数而言，其在每一个子区间上的函数都是线性的，但从整体上来看，分段线性函数表现为折线。其模型结构为：

$$g(t)=(k+\boldsymbol{a}(t)\boldsymbol{\delta})\cdot t+(m+\boldsymbol{a}(t)^T\boldsymbol{\gamma})$$

其中，k表示增长率，$\boldsymbol{\delta}$表示增长率的变化量，m为偏置参数。需要注意的是，γ的设置为$\boldsymbol{\gamma}=(\gamma_1,\ldots,\gamma_s)^T,\gamma_j=-s_j\delta_j$，这与逻辑回归函数中的设置有所区别。

Prophet 模型利用傅里叶级数来拟合时间序列的季节因素 s(t), 具体表述如下：

$$s(t)=\sum_{n=1}^{N}\left[a_n\cos(\frac{2\pi nt}{p})+b_n\sin(\frac{2\pi nt}{p})\right]$$

其中，p用来设置我们想要的时间序列的规则周期长度。当以天为单位时，对年度数据设置 p=365.25，周数据设置 p=7。这样，就得到了任意平滑周期效应的估算值$s(t)$。对于参数 N 的设定，一般年度数据设定 N=10, 周数据设定 N=3。

Prophet 模型通过设置虚拟变量，将不同节假日看成相互独立的模型，并可以通过设置不同的前后窗口值来表示节假日的影响范围，具体表述如下：

$$h(t) = \boldsymbol{Z}(t)\boldsymbol{\kappa} = \sum_{i=1}^{L} \kappa_i . 1_{\{t \in D_i\}}$$

其中，κ_i为节假日影响程度的大小（向量），$1_{\{t \in D_i\}}$为示性函数（向量），$\boldsymbol{Z}(t) = \left(1_{\{t \in D_1\}}, \ldots, 1_{\{t \in D_L\}}\right), \boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_L)^T$，$i$为第 i 个节假日，D_i 表示第 i 个节假日长度。

Prophet 模型就是通过拟合上述三项，然后最后把它们结合起来就得到了 Y 的最终拟合值。具体来说，Prophet 模型具体实施步骤主要包括如下六步。

第一步，观察序列 Y，了解该序列的基本特性并对异常点等进行预处理。

第二步，选择趋势模型。模型默认使用分段线性的趋势；如果依据先验知识，认为该预测对象表现为饱和增长或饱和减少，则可选择分段逻辑回归。

第三步，设置趋势转折点。如果存在已知特殊的转折点，即时间序列的趋势会在该位置发现转变，则可以进行人工设置。

第四步，设置周期性。模型默认是带有年和星期以及天的周期性，其他月、小时的周期性需要自己根据数据的特征进行设置，或者设置将年和星期等周期关闭。

第五步，设置节假日特征。依据对 Y 的了解，设置 holiday 参数来进行调节。可以设置不同的 holiday，影响大小不一样，时间段也不一样。

第六步，对各部分分别构建模型后，我们对其进行函数拟合，且需要根据模型是否过拟合以及对什么成分过拟合，来对应调节季节成分先验规模（seasonality_prior_scale）、节假日先验规模（holidays_prior_scale）、以及突变点先验规模（changepoint_prior_scale）等参数。

最后比较一下 Prophet 模型和 X-13 模型，并提出两个问题供读者思考。

两个模型的共同点是，都可以实现对 Y 的季节分解功能，从实现对 Y 的预测。两个模型的不同点在于，第一，X-13 模型是通过多次多种不同方法的反复移动平均来实现对 Y 的分解，而 Prophet 模型则是通过对不同因素分别建模来实现这种分解；第二，X-13 模型只能对月度数据、季度数据实现分解，而 Prophet 模型则除了上述频度数据之外，还可对日数据、周数据实现分解。

请读者思考如下两个问题：第一，X-13 模型中的 C 在 Prophet 模型中去哪儿了？而 Prophet 模型中的 h 在 X-13 模型中又怎么生？第二，对于月度数据，Prophet 模型中的参数 p 和 N 该如何确定？

主要参考文献

Taylor S. J, Letham B. Forecasting at scale [J]. The American Statistician, 2018, 72(1): 37-45.