Модели по данным разной частоты и их применения в задачах прогнозирования временных рядов

Цели работы:

- подготовка аналитического обзора моделей по смешанным данным;
- построение моделей по смешанным данным на реальных данных белорусской экономике;
- сравнительный анализ точности прогнозирования альтернативных типов моделей по смешанным данным.

Постановка задачи.

В качестве примера приложения моделей по данным разной частоты решается задача исследования зависимости показателя ВВП Беларуси от показателя ИПЦ (инфляции) Беларуси и обменных курсов валют относительно белорусского рубля.

Обычно все часто применяемые регрессионные модели машинного обучения работают с данными, заданными в одной частоте. Нередко на практике при анализе собранных данных можно столкнуться с такой проблемой, как различная частота этих данных. К примеру, некоторые данные из сферы экономики, как правило, формируются в квартальных представлениях. Параллельно с этим какие-либо объясняющие факторы могут быть собраны с более высокой частотой, будь то ежемесячные, еженедельные или ежедневные представления. Однако стандартные регрессионные модели не заточены под такое представление данных. Соответственно в ходе предварительного анализа необходимо преобразовать данные к одной частоте. В целях решения этой проблемы можно рассмотреть следующие подходы.

1. Одним из простейших вариантов решения рассматриваемой проблемы может оказаться наивное приведение данных более высокой частоты к нужной нам более низкой частоте, иначе говоря, агрегация данных более высокой частоты.

   Приведем пример: если исследуемая зависимая переменная находится в квартальном представлении, а независимые данные — в ежемесячном, то мы можем составить новый набор независимых переменных, взяв в качестве квартального значения последний месяц квартала.

   Однако такой подход имеет свой главный недостаток: возникает потеря некоторой информации о динамике объясняющих данных, которая может быть крайне полезна при построении модели.

2. Вторым вариантом сопоставления частот является интерполяция низкочастотных переменных. Для этого используются специальные подходы для заполнения пропущенных значений, рассматривать которые мы не будем. Этот вариант используется редко, и зачастую предпочтение отдается первому варианту.

   Этот подход также может способствовать появлению различного рода проблем при построении модели.

В связи с этим возникает вопрос: как можно без преобразования данных и потери какой-либо информации строить регрессионную модель для предсказания исследуемых показателей.

Одним из главных методов работы с данными смешанной частоты является mixed-data sampling метод, впервые представленный в работах Ghysels, Santa-Clara и Valkanov

(2004). MIDAS модели обрабатывают данные, отобранные с разной частотой, с использованием полиномов с распределенным запаздыванием. В то время как ранние исследования MIDAS были сосредоточены на финансовых приложениях, в последнее время этот метод используется для прогнозирования макроэкономических временных рядов, где обычно квартальный рост ВВП прогнозируется по ежемесячным макроэкономическим и финансовым показателям.

Совершенно другим методом работы с данными смешанной частоты являются векторные авторегресиионные модели (VAR), которые для предсказания используют не только прошлые значения объясняющих факторов, но и прошлые значения предсказываемой переменной. Таким образом, при прогнозировании они также будут учитывать поведение прогнозируемой переменной на рассматриваемом промежутке времени. К тому же, в отличие от MIDAS моделей, модели VAR также могут заполнять недостающие наблюдения для данных более низкой частоты.

Рассмотрение моделей начнем с простейшей модели с распределенным запаздыванием (distributed lag, DL), поскольку модели MIDAS регрессии имеют общие черты с этими моделями. Однако модели с распределенным запаздыванием проще по своей структуре.

Модель с распределенным запаздыванием, или DL-модель, может быть записана в следующем виде:

$$y_t = \beta_0 + \sum_{i=0}^{p} b_i x_{t-i} + \varepsilon_t, \tag{1}$$

Однако построенная модель может работать лишь с данными одной частоты, поэтому для использования этой модели нам надо агрегировать объясняющие показатели x_t, чтобы они имели одну частоту с прогнозируемым показателем y_t.

В качестве модификации можно рассматривать DL-модель с лагами Алмон.

Вообще говоря, DL-модель является частным случаем модели авторегрессии с распределенным запаздыванием (autoregressive distributed lag, ARDL). Ничто не мешает дополнить модель AR(p) некоторыми экзогенными переменными и их лагами, например, до порядка q. Такую модель называют открытой или же моделью авторегрессии с распределенным запаздыванием (autoregressive ditributed lags, ARDL(p, q))

$$\sum_{i=0}^{p} \beta_i y_{t-i} = \sum_{i=0}^{q} \alpha_j x_{t-j} + \varepsilon_t. \tag{2}$$

Но данная модель все также работает лишь с агрегированными данными.

Чтобы ввести базовую модель регрессии по данным смешанной частоты (mixed data sampling, MIDAS), изменим обозначения для переменных. Пусть эндогенная переменная y_t имеет фиксированную частоту. Она может быть годовая, квартальная, месячная и так далее. Для конкретики возьмем квартальную частоту. Кроме того, пусть независимая переменная замерена в m раз чаще. Например, если у эндогенной переменной квартальная частота, то для экзогенной переменной с месячной частотой возьмем $m = 3$.

Аналогично предыдущей DL-модели введем полиномиальный лаговый оператор следующего вида

$$b(L^{1/m}, \Theta) = \sum_{i=0}^{p} b(i, \Theta) L^{i/m}, \quad L^{i/m} x_t^{(m)} = x_{(t-i)/m}^{(m)}. \tag{3}$$

лючевую роль в результатах прогнозирования моделью MIDAS играет функция лаговых коэффициентов $b(i, \Theta), \quad i = 0, \ldots, p$. Ее можно задавать по-разному, что будет давать различные результаты. Фактически задание такой функции определяет способ

агрегации данных высокой частоты в ряд более низкой частоты (например, данные месячной частоты в данные квартальной частоты). Наиболее распространенными являются следующие виды функции лаговых коэффициентов:

- экспоненциальные лаги Алмон

$$b(i,\Theta) = \frac{e^{\Theta_1 i + \ldots \Theta_q i^q}}{\sum\limits_{j=0}^{p} e^{\Theta_1 j + \ldots \Theta_q j^q}}, \tag{4}$$

  где значение q либо задано априорно в самой программе, либо задается вручную;

- бета лаги (они требуют уже два параметра Θ)

$$b(i,\Theta_1,\Theta_2) = \frac{f(\frac{i}{p},\Theta_1;\Theta_2)}{\sum\limits_{j=0}^{p} f(\frac{j}{p},\Theta_1;\Theta_2)}, \quad f(x,\Theta_1,\Theta_2) = \frac{x^{a-1}(1-x)^{b-1}\Gamma(\Theta_1+\Theta_2)}{\Gamma(\Theta_1)\Gamma(\Theta_2)}; \tag{5}$$

В силу всех введенных обозначений, можем записать базовую модель MIDAS в следующем виде

$$y_t = \beta_0 + \beta_1 \cdot b(L^{1/m},\Theta)x_t^{(m)} + \varepsilon_t^{(m)}. \tag{6}$$

Также имеют место и другие модели предназначенные для работы по данным разной частоты:

- MIDAS-модели многих экзогенных переменных;
- нелинейные MIDAS-модели;
- многомерные MIDAS-модели;
- линейные MIDAS-модели с регуляризацией;
- U-MIDAS-модели, или неограниченные MIDAS-модели;
- MF-VAR, или векторная авторегрессия смешанной частоты;
- MF-BVAR, или байесовские векторные авторегрессии смешанной частоты;
- MS-MFVAR, или векторная авторегрессия по смешанным данным с марковскими переключениями состояний;
- DFM, или динамические факторные модели по смешанным данным.

Для оценки качества прогнозов моделей наиболее популярными являются три следующих критерия:

- средняя абсолютная ошибка (MAE);
- средняя абсолютная ошибка в процентах (MAPE);
- корень из среднеквадратической ошибки (RMSE).

При построении наукастов не учитывается информация о последнем доступном квартале: перед оцениванием модели из выборки удаляются значения зависимых переменных и соответствующие данному кварталу месячные значения объясняющих переменных. Далее в выборку возвращаются удаленные значения регрессоров и для них рассчитывается прогнозное значение зависимой переменной (наукаст). Рассматриваемые модели сравниваются по последним 12 точкам, в которых построены ретроспективные прогнозы, и проверяются на будущем прогнозе, который построен на невошедшем квартале.

Рассмотрим задачу для реальных данных. У нас имеется следующий набор данных

- эндогенная переменная — показатель внутреннего валового продукта (ВВП) Республики Беларусь на квартальной частоте;
- экзогенная переменная — показатели индекса потребительских цен (ИПЦ) Республики Беларусь на квартальной частоте и на месячной частоте;
- экзогенная (независимая) переменная — курс белорусского рубля к одной из валют: доллар, российский рубль — на дневной частоте.

Прежде чем строить модели, необходимо провести предварительный анализ и предобратботку переменных. Все построенные модели будут корректно работать только со стационарными временными рядами. Поэтому в первую очередь все временные ряды необходимо привести к стационарной форме. Параллельно с этим необходимо убрать сезонность и тренд (если они есть) в этих временных рядах.

Для всех MIDAS моделей выбиралось количество лагов для ежедневной экзогенной переменной равное 89, а для месячной экзогенной переменной 2. Для DL и ARDL моделей выбиралось число лагов для x_t равное 1, а в модели ARDL число лагов для y_t равное 4 (по результатам проверки эти значения лагов являются оптимальными).

Результаты оценки точности моделей представлены в таблицах 1, 2.

Таблица 1: Retrospective Evaluation Metrics

Model	MAE	MAPE	RMSE
MIDAS CPI_MM Beta	0.010561	0.473983	0.018825
MIDAS CPI_MM ExpAlmon	0.010561	0.473978	0.018825
MIDAS CPI_MM+RUB Beta	0.010875	0.816500	0.015378
MIDAS CPI_MM+RUB ExpAlmon	**0.010394**	**0.784645**	**0.013824**
MIDAS CPI_MM+USD Beta	0.013635	1.152350	0.016956
MIDAS CPI_MM+USD ExpAlmon	0.013467	0.993322	0.017887
DL CPI_QQ	0.016318	1.001932	0.023231
ARDL CPI_QQ	**0.012386**	**1.122417**	**0.015147**

Таблица 2: Future Evaluation Metrics

Model	MAE	MAPE	RMSE
MIDAS CPI_MM Beta	0.004407	0.787724	0.004407
MIDAS CPI_MM ExpAlmon	0.004407	0.787740	0.004407
MIDAS CPI_MM+RUB Beta	**0.001626**	**0.290612**	**0.006263**
MIDAS CPI_MM+RUB ExpAlmon	**0.001587**	**0.283747**	**0.006301**
MIDAS CPI_MM+USD Beta	0.004614	0.824704	0.012502
MIDAS CPI_MM+USD ExpAlmon	0.002240	0.400436	0.005648