SAS/ETS: модели прогнозирования

ЛЕКЦИЯ 2

Валентина Власова
Valentina.Vlasova@sas.com

Содержание

SAS/ETS

- Регрессионные модели и методы выбора параметров
  - Метод наименьших квадратов
  - Метод максимального правдоподобия
- Модели авторегрессии и скользящего среднего
  - Модель скользящего среднего
  - Авторегрессионная модель
  - ARMA
  - ARIMA
- Модели экспоненциального сглаживания
- Модели ненаблюдаемых компонент
- Векторная авторегрессия

РЕГРЕССИОННЫЕ МОДЕЛИ И МЕТОДЫ ВЫБОРА ПАРАМЕТРОВ

МОДЕЛЬ ПАРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$y_t = a + bx_t + \epsilon_t$$

где y_t – зависимая переменная,

x_t – объясняющая переменная,

$a+bx_t$ – неслучайная составляющая,

a, b – параметры уравнения,

ε_t – случайная составляющая.

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Пусть: x – набор из m неизвестных параметров,
$f_i(x)$ – функции от этого набора параметров,
y_i – значения выборки, которые приближаются с помощью функции $f_i(x)$

Метод наименьших квадратов заключается в выборе в качестве «меры близости» суммы квадратов отклонений значений функции от значений выборки:

$$|f_i(x) - y_i|$$

Таким образом, для нахождения параметров по МНК необходимо решить следующую задачу:

$$\sum_i e_i^2 = \sum_i (y_i - f_i(x))^2 \to \min_x$$

МНК ДЛЯ ЛИНЕЙНОЙ РЕГРЕССИИ

Пусть y – вектор-столбец наблюдений объясняемой переменной,
X – это (*n* x *k*)-матрица наблюдений факторов
(строки матрицы — векторы значений факторов в данном наблюдении,
по столбцам — вектор значений данного фактора во всех наблюдениях)

Матричное представление линейной модели имеет вид:

$$y = Xb + \varepsilon$$

Тогда вектор оценок объясняемой переменной и вектор остатков регрессии будут равны:

$$\hat{y} = Xb, \quad e = y - \hat{y} = y - Xb$$

Соответственно сумма квадратов остатков регрессии будет равна:

$$RSS = e^T e = (y - Xb)^T (y - Xb)$$

МНК ДЛЯ ЛИНЕЙНОЙ РЕГРЕССИИ

Дифференцируя эту функцию по вектору параметров *b* и приравняв производные к нулю, получим систему уравнений (в матричной форме):

$$(X^T X)b = X^T y$$

В расшифрованной матричной форме эта система уравнений выглядит следующим образом:

$$\begin{pmatrix} \sum x_{t1}^2 & \sum x_{t1}x_{t2} & \sum x_{t1}x_{t3} & \dots & \sum x_{t1}x_{tk} \\ \sum x_{t2}x_{t1} & \sum x_{t2}^2 & \sum x_{t2}x_{t3} & \dots & \sum x_{t2}x_{tk} \\ \sum x_{t3}x_{t1} & \sum x_{t3}x_{t2} & \sum x_{t3}^2 & \dots & \sum x_{t3}x_{tk} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{tk}x_{t1} & \sum x_{tk}x_{t2} & \sum x_{tk}x_{t3} & \dots & \sum x_{tk}^2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} \sum x_{t1}y_t \\ \sum x_{t2}y_t \\ \sum x_{t3}y_t \\ \vdots \\ \sum x_{tk}y_t \end{pmatrix},$$

где все суммы берутся по всем допустимым значениям *t*

МНК ДЛЯ ЛИНЕЙНОЙ РЕГРЕССИИ

Решение этой системы уравнений и дает общую формулу МНК-оценок для линейной модели:

$$\hat{b}_{OLS} = (X^T X)^{-1} X^T y = \left(\frac{1}{n} X^T X\right)^{-1} \frac{1}{n} X^T y = V_x^{-1} C_{xy}$$

В случае парной линейной регрессии, система уравнений имеет вид:

$$\begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix}$$

Отсюда несложно найти оценки коэффициентов:

$$\begin{cases} \hat{b} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}, \\ \hat{a} = \bar{y} - b\bar{x}. \end{cases}$$

МЕТОД МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

Пусть есть выборка $X_1, \ldots, X_n$ из распределения $\mathbb{P}_\theta$,
где $\theta \in \Theta$ – неизвестные параметры.

Пусть $L(\mathbf{x} \mid \theta) \colon \Theta \to \mathbb{R}$ – функция правдоподобия, где $\mathbf{x} \in \mathbb{R}^n$

Точечная оценка

$$\hat{\theta}_{\text{МП}} = \hat{\theta}_{\text{МП}}(X_1, \ldots, X_n) = \arg\max_{\theta \in \Theta} L(X_1, \ldots, X_n \mid \theta)$$

называется **оценкой максимального правдоподобия** параметра θ.

Оценка максимального правдоподобия – это такая оценка, которая максимизирует функцию правдоподобия при фиксированной реализации выборки.

МОДЕЛИ АВТОРЕГРЕССИИ И СКОЛЬЗЯЩЕГО СРЕДНЕГО

МОДЕЛЬ СКОЛЬЗЯЩЕГО СРЕДНЕГО

Модель скользящего среднего *q*-го порядка ***MA(q)***:

$$X_t = \sum_{j=0}^{q} b_j \varepsilon_{t-j}$$

ε_t – нормальный белый шум – последовательность независимых и одинаково распределённых по нормальному закону случайных величин, с нулевым средним и дисперсией σ^2;

b_j – параметры модели.

Эта модель содержит $q+1$ параметр (b_1, b_2, …, b_q и σ), значения которых нужно оценить по временному ряду.

МОДЕЛЬ СКОЛЬЗЯЩЕГО СРЕДНЕГО: МЕТОДЫ ОЦЕНКИ

1. МНК не подходит (сумма квадратов остатков не выражается аналитически)
2. Можно использовать ММП в предположении нормальности распределения
3. Альтернативный подход (асимптотически эквивалентный ММП):

Если предположить, что в периоды до наших наблюдений (до момента, с которого имеются данные по временному ряду) значения равны нулю, то получим:

$$x_1 = \varepsilon_1 \ , \ x_2 = \varepsilon_2 + b_1\varepsilon_1 \ , \ x_3 = \varepsilon_3 + b_1\varepsilon_2 + b_2\varepsilon_1, ...$$

Следовательно, в качестве остатков можно использовать последовательные выражения:

$$e_1 = x_1 \ , \ e_2 = x_2 - b_1 e_1 \ , \ e_3 = x_3 - b_1 e_2 - b_2 e_1, ...$$

Минимизируя сумму квадратов этих остатков по параметрам получим требуемые оценки.

АВТОРЕГРЕССИОННАЯ МОДЕЛЬ

Авторегрессионная (**AR-**) **модель** — модель временных рядов, в которой значения временного ряда в данный момент линейно зависят от предыдущих значений этого же ряда. Авторегрессионный процесс порядка *p* (AR(*p*)-процесс) определяется следующим образом:

$$X_t = c + \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t,$$

где $\alpha_1, \ldots, \alpha_p$ — параметры модели,

c — константа,

$\{\varepsilon_t\}$ — белый шум.

Эта модель содержит p+1 параметр (a1, a2, …, ap и σ), значения которых нужно оценить по временному ряду.

АВТОРЕГРЕССИОННОЕ СКОЛЬЗЯЩЕЕ СРЕДНЕЕ

Модель ARMA состоит из двух частей:

- авторегрессионная (AR)
- скользящее среднее (MA)

Обозначение модели ARMA(*p*,*q*), где *p* — порядок регрессионной части, а *q* — порядок скользящего среднего:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i}.$$

где

c — константа,

$\{\varepsilon_t\}$ — белый шум,

$\alpha_1, \ldots, \alpha_p$, $\beta_1, \ldots, \beta_q$ — действительные числа, авторегрессионные коэффициенты и коэффициенты скользящего среднего, соответственно.

АВТОРЕГРЕССИОННОЕ ИНТЕГРИРОВАННОЕ СКОЛЬЗЯЩЕЕ СРЕДНЕЕ

Модель **ARIMA(*p,d,q*)**, где *p,d* и *q* — целые неотрицательные числа, характеризующие порядок для частей модели (соответственно авторегрессионной, интегрированной и скользящего среднего):

$$\triangle^d X_t = c + \sum_{i=1}^{p} a_i \triangle^d X_{t-i} + \sum_{j=1}^{q} b_j \varepsilon_{t-j} + \varepsilon_t$$

ARIMA – расширение моделей ARMA для нестационарных временных рядов, которые можно сделать стационарными взятием разностей некоторого порядка от исходного временного ряда.
Модель *ARIMA(p, d, q)* означает, что разности временного ряда порядка *d* подчиняются модели *ARMA(p, q).* То есть при *d=0* получаем обычные *ARMA*-модели.

SAS/ETS | МОДЕЛИ ARIMA

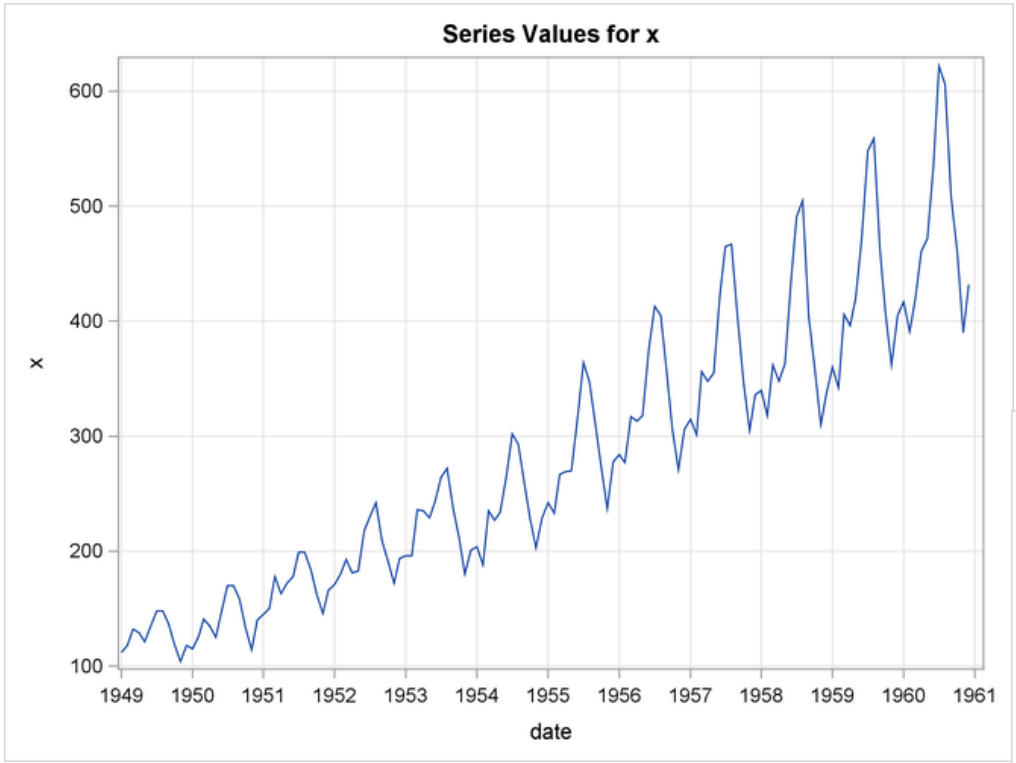
Процедура ARIMA позволяет проанализировать и спрогнозировать значения одномерных временных рядов с помощью моделей авторегрессии интегрированного скользящего среднего (ARIMA) или авторегрессии скользящего среднего (ARMA).

Модель ARIMA предсказывает значение временного ряда в виде линейной комбинации собственных прошлых значений ряды, прошлых ошибок и текущих и прошлых значений других временных рядов.

PROC ARIMA options ;
BY variables ;
IDENTIFY VAR=variable options ;
ESTIMATE options ;
OUTLIER options ;
FORECAST options ;

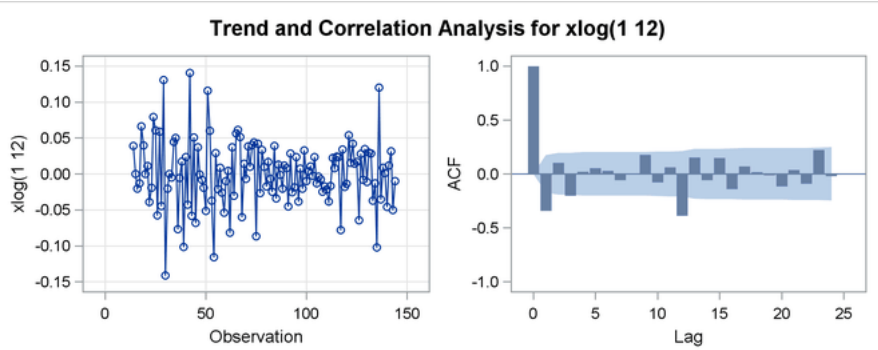
1. Анализ временного ряда и определение возможных моделей прогнозирования
2. Оценка параметром моделей
3. Построение прогноза с помощью настроенной модели

МОДЕЛИ ARIMA: ПРИМЕР



Модель ARIMA $(0, 1, 1)x(0, 1, 1)_{12}$

```
proc arima data=seriesg;
   identify var=xlog(1,12);
   estimate q=(1)(12) noint method=ml;
   forecast id=date interval=month
printall out=b;
run;
```



МОДЕЛИ ЭКСПОНЕНЦИАЛЬНОГО СГЛАЖИВАНИЯ

ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ

Пусть задан временной ряд: $X = \{x_1, \cdots x_T\}$

Экспоненциальное сглаживание ряда осуществляется по рекуррентной формуле:

$$S_t = \alpha x_t + (1-\alpha) S_{t-1}, \quad \alpha \in (0,1)$$

Для прогнозирования используется следующая модель (модель Брауна):

$$\hat{y}_{t+d} = \alpha y_t + (1-\alpha)\hat{y}_t, \ \hat{y}_0 = y_0, \ \alpha \in (0,1)$$

Для более быстрого отражения новых изменений следует увеличивать вес последних наблюдений: $\alpha \to 1$.

Для сглаживания случайных отклонений α нужно уменьшать: $\alpha \to 0$.

Эмпирические правила: если $\alpha \in (0, 0.3)$, то ряд стационарен, модель работает; если $\alpha \in (0.3, 1)$, то ряд нестационарен и нужна трендовая модель.

ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ

Модель Брауна работает только при небольшом горизонте прогнозирования, так как не учитываются тренд и сезонные изменения!

Чтобы учесть их влияние, можно использовать следующие модели:

- Модель Хольта – учитывается линейный тренд;
- Хольта-Уинтерса – мультипликативные экспоненциальный тренд и сезонность;
- Тейла-Вейджа – аддетивные линейный тренд и сезонность:

$$
\begin{aligned}
\hat{y}_{t+d} &= a_t + d b_t \circledcirc_{t+(d \bmod s)-s}, \\
a_t &= \alpha_1 \left(y_t - \circledcirc_{t-s}\right) + \left(1-\alpha_1\right)\left(a_{t-1} + b_{t-1}\right), \\
b_t &= \alpha_3 (a_t - a_{t-1}) + \left(1-\alpha_3\right) b_{t-1}, \\
\circledcirc_t &= \alpha_2 \left(y_t - a_t\right) + \left(1-\alpha_2\right) \circledcirc_{t-s},
\end{aligned}
$$

SAS/ETS ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ

Процедура ESM строит прогнозы с помощью моделей экспоненциального сглаживания с оптимальными весами сглаживания.

Позволяет использовать следующие модели сглаживания:

- простое сглаживание
- двойное сглаживание
- линейное сглаживание
- модель с затухающим трендом
- сезонная модель сглаживания
- метод Уинтерса (аддитивный и мультипликативный)

PROC ESM options ;

BY variables ;

ID variable INTERVAL= interval options ;

FORECAST variable-list / options ;

ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ И ARIMA

Экспоненциальное сглаживание ряда осуществляется по рекуррентной формуле:

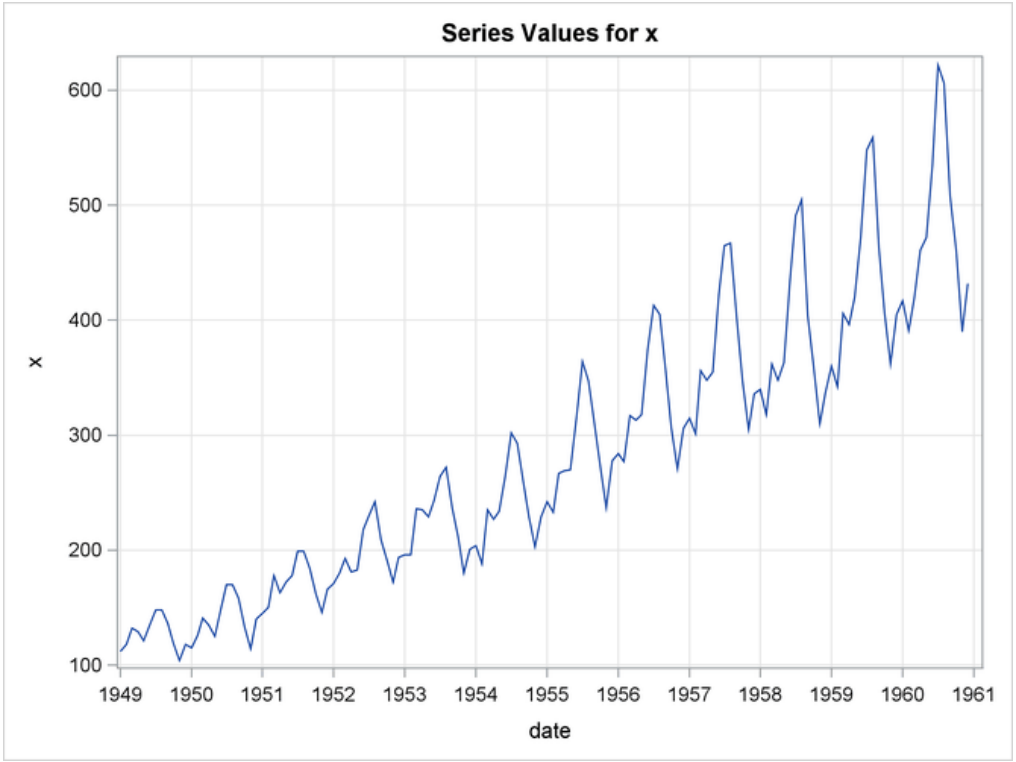
$$S_t = \alpha x_t + (1-\alpha) S_{t-1}, \quad \alpha \in (0,1)$$

Если последовательно использовать это рекуррентное соотношение, то можно получить следующую формулу:

$$S_t = \alpha x_t + (1-\alpha)\left(\alpha x_{t-1} + (1-\alpha) S_{t-2}\right) = \cdots = \alpha \sum_{i=0}^{t-1} (1-\alpha)^i x_{t-i} + (1-\alpha)^t S_0$$

То есть тоже может быть описано с помощью моделей ARIMA.

SAS/ETS | ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ



- Тренд, возможно экспоненциальный
- Мультипликативная сезонность

МОДЕЛИ НЕНАБЛЮДАЕМЫХ КОМПОНЕНТ

МОДЕЛЬ НЕНАБЛЮДАЕМЫХ КОМПОНЕНТ

Модель ненаблюдаемых компонент можно рассматривать как модель множественной регрессии с изменяющимися во времени коэффициентами. Она основана на следующих принципах:

- Временной ряд может быть разложен на трендовую, сезонную и циклическую компоненты.
- Модели временных рядов, которые дают равный вес ближним и отдаленным наблюдениям, неприменимы.

МОДЕЛЬ НЕНАБЛЮДАЕМЫХ КОМПОНЕНТ

Модель ненаблюдаемых компонент раскладывает временной ряд на компоненты, такие как тренд, сезонность, циклические составляющие и регрессионные эффекты, связанные с объясняющей переменной. Следующая функция показывает возможный вариант модели:

$$y_t = \mu_t + \gamma_t + \psi_t + \sum_{j=1}^{m} \beta_j x_{jt} + \varepsilon_t$$

$$\varepsilon_t \sim i.i.d.\ N(0, \sigma_\varepsilon^2)$$

Где μ_t, γ_t и ψ_t представляют тренд, сезонность и циклическую компоненту соответственно.

ФУНКЦИИ ТРЕНДА

Процедура UCM предлагает два способа для моделирования трендовой составляющей:

- Модель случайного блуждания (подразумевается, что тренд остается примерно постоянным в течение всего временного ряда без постоянного снижения или роста):

$$\mu_t = \mu_{t-1} + \eta_t, \quad \eta_t \sim i.i.d.\ N(0, \sigma_\eta^2)$$

- Во второй модели тренд моделируется как локально линейнейная функция, состоящая из уровня и наклона:

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t, \quad \eta_t \sim i.i.d.\ N(0, \sigma_\eta^2)$$
$$\beta_t = \beta_{t-1} + \xi_t, \quad \xi_t \sim i.i.d.\ N(0, \sigma_\xi^2)$$

МОДЕЛИРОВАНИЕ ЦИКЛА

- Детерминированный цикл ψ_t с частотой $\lambda, 0 < \lambda < \pi$ может быть описан следующей функцией:

$$\psi_t = \alpha \cos(\lambda t) + \beta \sin(\lambda t)$$

Если аргумент *t* измеряется на непрерывной шкале, то ψ_t периодическая функция с периодом $2\pi/\lambda$, амплитудой $\gamma = (\alpha^2 + \beta^2)^{1/2}$ и фазой $\varphi = \tan^{-1}(\beta/\alpha)$. Уравнение цикла может быть записана в терминах амплитуды и фазы следующим образом:

$$\psi_t = \gamma \cos(\lambda t - \varphi)$$

МОДЕЛИРОВАНИЕ СЕЗОННОСТИ

Сезонные колебания являются распространенным источником изменений в данных временных рядов.

Эти колебания возникают из-за регулярных изменений в разные сезоны или некоторых других периодических событий.

Сезонные эффекты рассматривают как поправки к общему тренду ряда из-за сезонных колебаний, сумма этих эффектов равна нулю при суммировании по полному циклу.

$$\sum_{i=0}^{s-1} \gamma_{t-i} = \omega_t, \qquad \omega_t \sim i.i.d.\ N(0, \sigma_\omega^2)$$

Для формирования прогноза с использованием модели ненаблюдаемых компонент используется процедура UCM.

```
PROC UCM <options> ;
   AUTOREG <options> ;
   BLOCKSEASON options ;
   BY variables ;
   CYCLE <options> ;
   DEPLAG options ;
   ESTIMATE <options> ;
   FORECAST <options> ;
   ID variable options ;
   IRREGULAR <options> ;
   LEVEL <options> ;
   MODEL dependent variable <= regressors> ;
   NLOPTIONS options ;
   PERFORMANCE options ;
   OUTLIER options ;
   RANDOMREG regressors </ options> ;
   SEASON options ;
   SLOPE <options> ;
   SPLINEREG regressor <options> ;
   SPLINESEASON options ;
```

ВЕКТОРНАЯ АВТОРЕГРЕССИЯ

ВЕКТОРНАЯ АВТОРЕГРЕССИЯ

Векторная авторегрессия (*VAR, Vector AutoRegression*) – модель динамики нескольких временных рядов, в которой текущие значения этих рядов зависят от прошлых значений этих же временных рядов.

Фактически VAR – это система эконометрических уравнений, каждая из которых представляет собой модель ADL.

Пусть y^i, *1,…,k* – *i*-й временной ряд. *ADL(p,p)*-модель для *i*-го временного ряда:

$$y_t^i = a_0^i + \sum_{j=1}^{k} a_{1j}^i y_{t-1}^j + \sum_{j=1}^{k} a_{2j}^i y_{t-2}^j + \ldots + \sum_{j=1}^{k} a_{pj}^i y_{t-p}^j + \varepsilon_t^i$$

ВЕКТОРНАЯ АВТОРЕГРЕССИЯ

Если ввести вектор временных рядов $y_t = (y_t^1, y_t^2, \ldots, y_t^k)$ и матрицы коэффициентов $A_m = \{a_{mj}^i\}$, тогда *ADL* уравнения для каждого временного ряда можно записать одним уравнением в векторной форме:

$$y_t = a_0 + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + \varepsilon_t = a_0 + \sum_{m=1}^{p} A_m y_{t-m} + \varepsilon_t$$

Такая модель является *замкнутой*, в том смысле, что в качестве объясняющих переменных выступают только лаги эндогенных (объясняемых) переменных.

Если дополнить модель экзогенными переменными и их лагами, то получим модель называемую *открытой*:

$$y_t = a_0 + \sum_{m=1}^{p} A_m y_{t-m} + \sum_{n=0}^{q} B_n x_{t-n} + \varepsilon_t$$

ВЕКТОРНАЯ АВТОРЕГРЕССИЯ

Процедура VARMAX позволяет смоделировать динамическую взаимосвязь между несколькими взаимосвязанными зависимыми, а также независимыми временными рядами:

```
PROC VARMAX  options ;
        BOUND  restriction, …, restriction ;
        BY  variables ;
        CAUSAL  GROUP1=(variables) GROUP2=(variables) ;
        COINTEG  RANK=number <options> ;
        GARCH  options ;
        ID  variable INTERVAL=value <ALIGN=value> ;
        INITIAL  equation, …, equation ;
        MODEL  dependents < = regressors > <, dependents < = regressors > …> < / options > ;
        NLOPTIONS  options ;
        OUTPUT  <options> ;
        RESTRICT  restriction, …, restriction ;
        TEST  restriction, …, restriction ;
```

ВЕКТОРНАЯ АВТОРЕГРЕССИЯ

Рассмотрим простейшую модель:

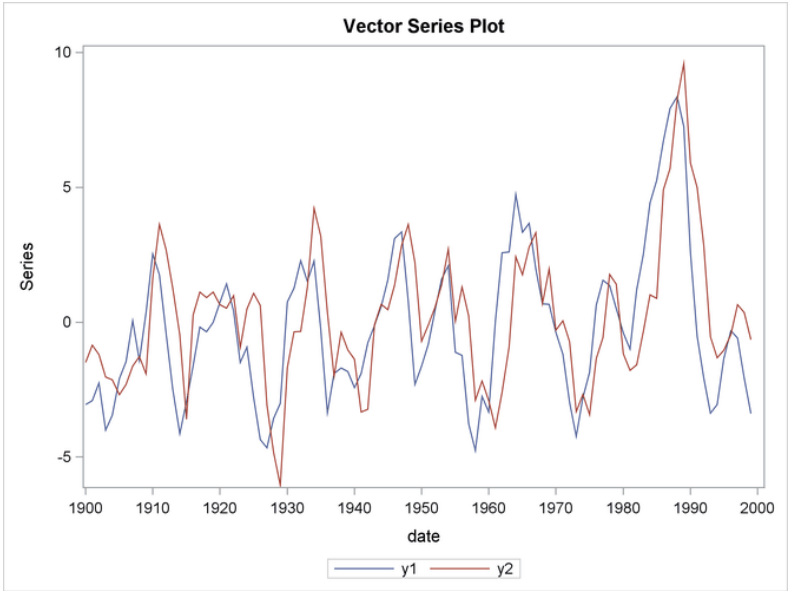
$$\mathbf{y}_t = \begin{pmatrix} 1.2 & -0.5 \\ 0.6 & 0.3 \end{pmatrix} \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \text{with } \Sigma = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.25 \end{pmatrix}$$

Смоделируем данные:

```
proc iml;
          sig = {1.0 0.5, 0.5 1.25};
          phi = {1.2 -0.5, 0.6 0.3}; /* simulate the vector time series */
          call varmasim(y,phi) sigma = sig n = 100 seed = 34657;
          cn = {'y1' 'y2'};
          create simul1 from y[colname=cn];
          append from y;
quit;
```

И построим график:

```
data simul1;
          set simul1;
          date = intnx( 'year', '01jan1900'd, _n_-1 );
          format date year4.;
run;
ods graphics on;
proc timeseries data=simul1 vectorplot=series;
          id date interval=year;
          var y1 y2;
run;
```



ВЕКТОРНАЯ АВТОРЕГРЕССИЯ

Процедура для моделирования и прогнозирования будет выглядеть так:

```
proc varmax data=simul1;
          id date interval=year;
          model y1 y2 / p=1 noint lagmax=3 print=(estimates diagnose);
          output out=for lead=5;
run;
```

В результате получим следующую модель:

$$\mathbf{y}_t = \begin{pmatrix} 1.160 & -0.511 \\ (0.055) & (0.059) \\ 0.546 & 0.385 \\ (0.058) & (0.062) \end{pmatrix} \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t$$

Если записать иначе:

$$\begin{aligned} y_{1t} &= 1.160\, y_{1,t-1} - 0.511\, y_{2,t-1} + \varepsilon_{1t} \\ y_{2t} &= 0.546\, y_{1,t-1} + 0.385\, y_{2,t-1} + \varepsilon_{2t} \end{aligned}$$

Спасибо за внимание!

SAS.com