Временные ряды

Временной ряд (ВР) – это данные, характеризующие один и тот же объект в различные моменты времени (временной срез)

$$y_1, y_2, \ldots, y_n \text{ или } y_t \quad (t = \overline{1,n}).$$

$y_t \quad (t = \overline{1,n})$ – ***уровни*** (***элементы***) ряда, где n – число уровней.

Каждый уровень ВР формируется под воздействием факторов, которые условно подразделяются на составляющие:

- факторы, формирующие тенденцию (тренд) ряда;
- факторы, формирующие периодические колебания ряда;
- случайные факторы.

Тренд или ***тенденция*** ***T*** ВР – это систематическая (неслучайная) составляющая, определяющее общее направление развития ВР.

Наряду с тенденциями во ВР часто имеют место более или менее регулярные колебания – ***периодические*** составляющие.

Если период колебаний ≤ 1 года, то их называют ***сезонными*** ***S***.

При большем периоде колебания – имеет место ***циклическая*** составляющая ***C***.

Циклические изменения *не обязательно периодические*.

Если из ВР удалить тренд и периодические составляющие, то останется ***нерегулярная*** (***случайная***) ***компонента*** *U*.

Модель, в которой ВР представлен в виде суммы соответствующих компонент, называется ***аддитивной.*** Модель, в которой ВР представлен в виде произведения – ***мультипликативной***:

аддитивные модели ряда

$$y_t = T + C + S + U,$$

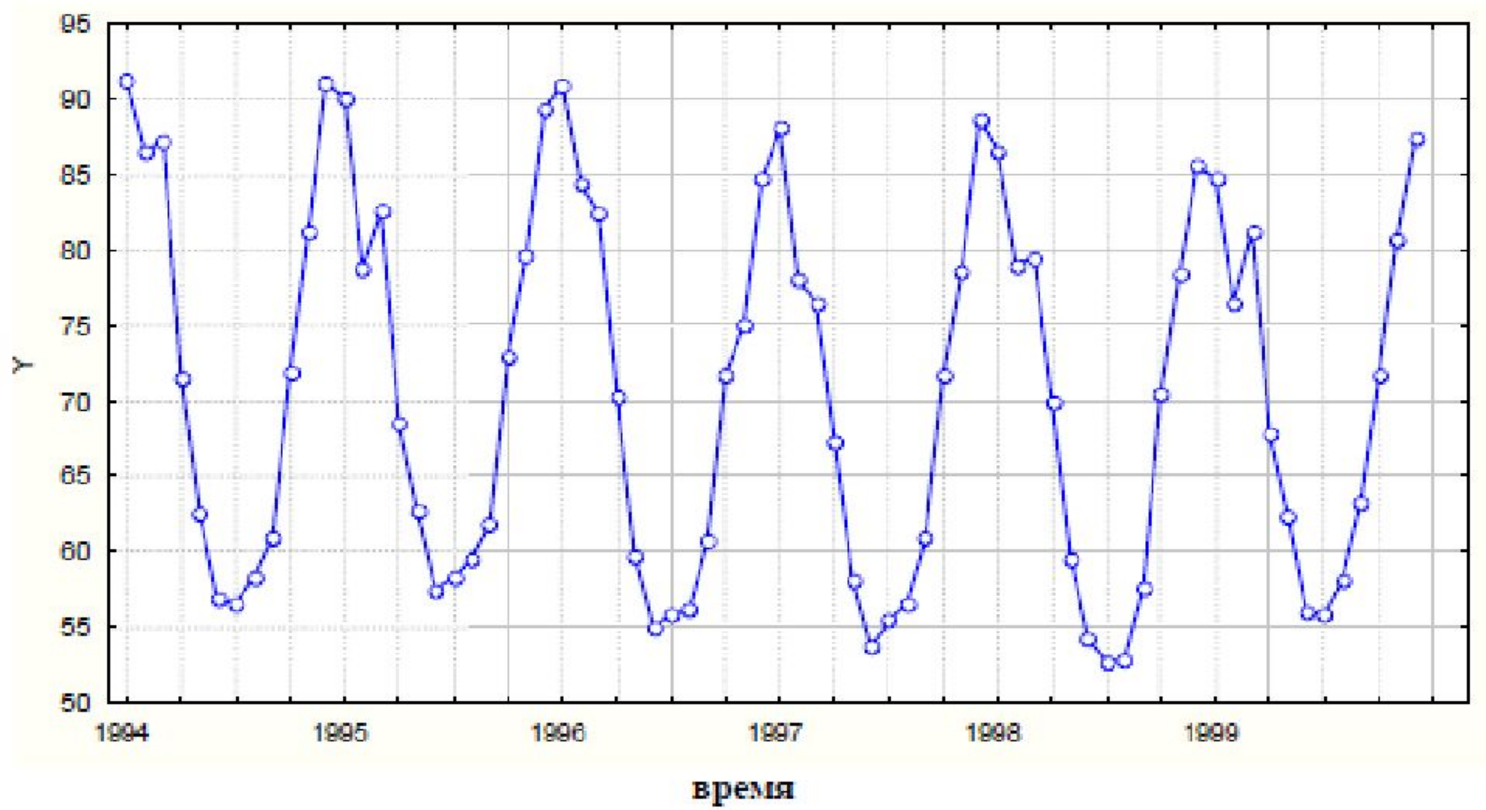
мультипликативные модели

$$y_t = T \cdot C \cdot S \cdot U.$$

Смешанные модели

$$y_t = T \cdot C \cdot S + U$$

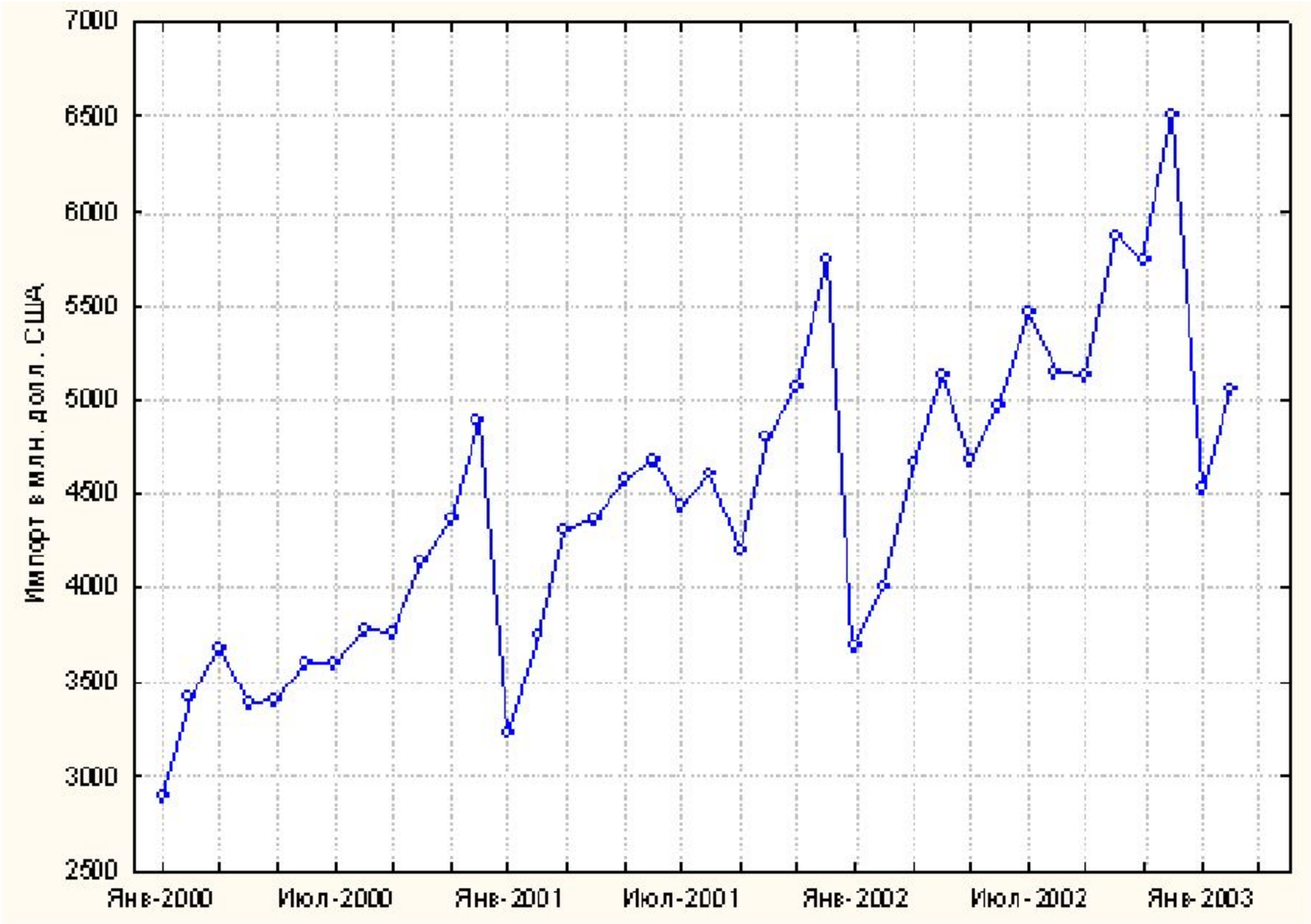
Решение задачи по ВР начинается с построения *графика*.



Динамика производства электроэнергии в РФ (млн. кВт.ч)

Тенденция этого ряда была близка к линейному развитию.

Видны устойчивые сезонные колебания, имеющие годовую периодичность.



Временной ряд объема импорта РФ.

Заметен растущий тренд.

Ярко выражена сезонная составляющая с минимумом в январе и максимумом в декабре.

ИССЛЕДОВАНИЕ ТЕНДЕНЦИИ

ВРЕМЕННОГО РЯДА

Проверка гипотезы о существовании тренда.

Для выявления факта наличия или отсутствия неслучайной составляющей, т.е. для проверки гипотезы о независимости средней ВР от времени:

$$H_o : M(y(t)) = a = const,$$

используют следующие критерии.

Метод сравнения средних.

ВР разбивают на 2 примерно равные части n_1 и n_2: $Y_1 = (y_1, \ldots y_{n_1})$ и $Y_2 = (y_{n_1+1}, \ldots y_{n_2})$, каждая из которых самостоятельная выборочная совокупность, имеющие нормальное распределение.

Решаются две задачи.

Если ВР имеет тенденцию, то $\overline{Y_1}$ и $\overline{Y_2}$ должны существенно различаться между собой.

Проверяется гипотеза о равенстве средних:

$$H_0 : M(Y_1) = M(Y_2) .$$

За основу проверки берется t-критерий Стьюдента.

В случае равенства дисперсий ($\sigma_1^2 = \sigma_2^2$):

$$t_{набл} = \frac{\overline{Y_1} - \overline{Y_2}}{S_{\overline{Y_1}-\overline{Y_2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(\alpha, v = n_1 + n_2 - 2), \text{ где}$$

$S_{\overline{Y_1}-\overline{Y_2}}$ – выборочное среднеквадратическое отклонение:

$$S_{\overline{Y_1}-\overline{Y_2}} = \sqrt{\frac{(n_1 - 1)^2 \cdot S_1^2 + (n_2 - 1)^2 \cdot S_2^2}{n_1 + n_2 - 2}}, \text{ где}$$

$$S_1^2 = \frac{\sum_{i=1}^{n_1} (y_i - \overline{Y_1})^2}{n_1 - 1} \quad \text{и} \quad S_2^2 = \frac{\sum_{i=n_1+1}^{n} (y_i - \overline{Y_2})^2}{n_2 - 1}$$

Проверка гипотезы о равенстве дисперсий

$H_0 : \sigma_1^2 = \sigma_2^2; \; H_1 : \sigma_1^2 \neq \sigma_2^2$

осуществляется с помощью F-критерия:

$$F_{расч} = \frac{s_2^2}{s_1^2},$$

где дисперсии для 1–ой и 2–ой частей ряда рассчитываются по формуле:

$$s_i^2 = \frac{\sum (y_i - \overline{y_i})^2}{n-1}$$

Если $F_{расч} \le F_{крит}$, то гипотеза о равенстве дисперсий принимается. Если $F_{расч} > F_{крит}$, то гипотеза о равенстве дисперсий отклоняется.

АВТОКОРРЕЛЯЦИЯ УРОВНЕЙ ВРЕМЕННОГО РЯДА.

При наличии во ВР тенденции и периодических колебаний значения каждого y_t зависят от $y_1, y_2, \ldots, y_{t-1}$.

Корреляционная зависимость между последовательными уровнями ВР называется ***автокорреляцией уровней***.

Степень тесноты корреляционной связи между y_t может быть определена с помощью *коэффициентов автокорреляции*, т.е. *линейным коэффициентом корреляции* между $y_1, y_2, \ldots, y_t, \ldots, y_n$ и $(y_{1+\tau}, y_{2+\tau}, \ldots, y_{t+\tau}, \ldots, y_{n+\tau})$.

Формула для расчета коэффициента корреляции имеет вид:

$$r(\tau)=\frac{M[(y_t-a_{y_t})(y_{t+\tau}-a_{y_{t+\tau}})]}{\sqrt{D(y_t)D(y_{t+\tau})}}=\frac{\operatorname{cov}(y_t,y_{t+\tau})}{\sigma(y_t)\sigma(y_{t+\tau})}.$$

Число периодов, по которым рассчитывается коэффициент автокорреляции, называют ***лагом***.

Выборочной оценкой $r(\tau)$ является ***выборочный коэффициент автокорреляции*** r_τ:

$$r_\tau=\frac{\operatorname{cov}(y_t,y_{t+\tau})}{\sigma(y_t)\sigma(y_{t+\tau})}$$

Свойства коэффициента автокорреляции.

1. r_τ характеризует тесноту только линейной связи текущего и предыдущего уровней ряда.

2. Для ВР, имеющего сильную нелинейную тенденцию, r_τ может приближаться к нулю.

3. По знаку r_τ нельзя делать вывод о возрастающей или убывающей тенденции в уровнях ряда. Большинство ВР экономических данных содержат положительную автокорреляцию уровней, однако при этом могут иметь убывающую тенденцию.

Порядок коэффициентов автокорреляции определяется временным лагом.

Последовательность коэффициентов автокорреляции $r(1)$, $r(2)$ и т.д. называют ***автокорреляционной функцией*** ВР (АКФ).

График зависимости ее значений от величины лага называется ***коррелограммой***.

При помощи АКФ можно выявить структуру ВР:

- если наиболее высоким оказался $r(1)$, то исследуемый ряд содержит только ***линейную тенденцию***.
- если наиболее высоким оказался $r(\tau)$, то ряд ***содержит циклические колебания с периодичностью в*** τ моментов времени.
- если ни один из коэффициентов автокорреляции не является значимым, то либо ряд не содержит тенденции и циклических колебаний, либо ряд содержит сильную нелинейную тенденцию.

Прогнозирование временных рядов

с помощью кривых роста

Одним из приемов изучения общей тенденции во ВР является ***аналитическое выравнивание***.

Аналитическим выравниванием ВР называют нахождение аналитической функции времени $\hat{y}=f(t)$, характеризующей основную тенденцию изменения уровней ряда с течением времени.

Функция $f(t)$ носит название ***кривой роста***.

Основные виды кривых

Кривая роста позволяет получить *выровненные* или *теоретические* значения уровней ВР - $\hat{y}_t$ (это уровни, которые наблюдались бы в случае полного совпадения динамики явления с кривой).

Процедура разработки прогноза с использованием кривых роста включает в себя следующие этапы:

1) выбор одной или нескольких кривых, форма которых соответствует характеру изменения ВР;

2) оценка параметров выбранных кривых;

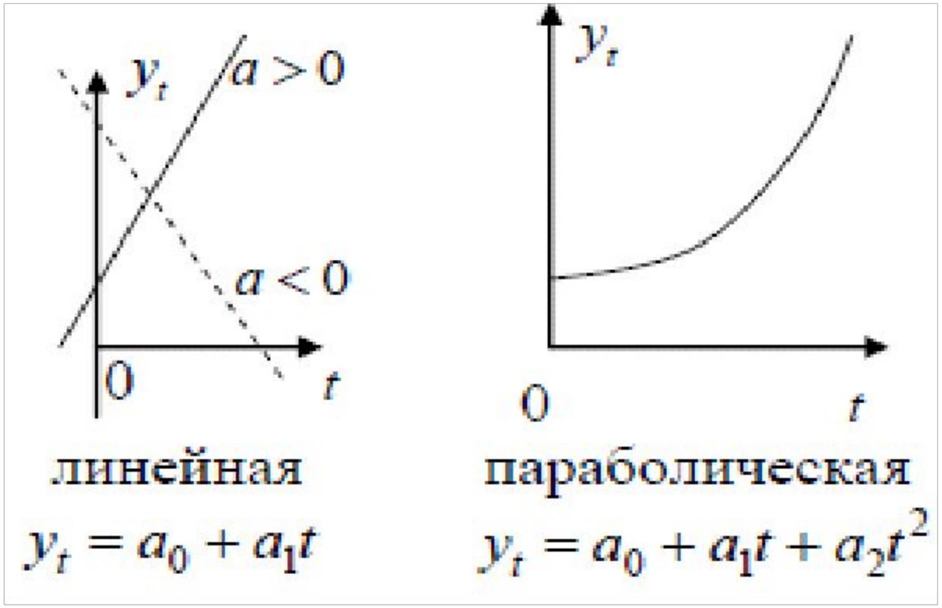
3) проверка адекватности выбранных кривых и окончательный выбор кривой роста;

4) расчет точечного и интервального прогнозов.

Наиболее часто на практике используются кривые роста трех основных типов.

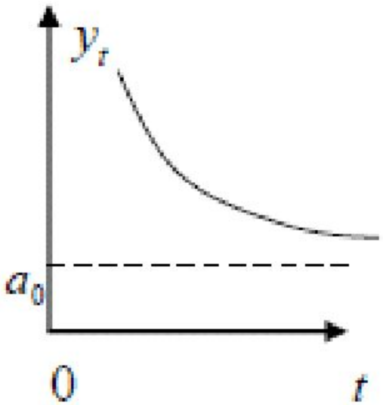
I класс – функции, используемые для описания процессов *с монотонным характером развития без предела роста*.

- ***полиномы*** $\hat{y}_t = a_0 + a_1 t + a_2 t^2 + \ldots + a_p t^p$;
- ***показательная*** - $\hat{y}_t = ab^t$, где b – темп изменения;
- ***линейно–логарифмическая*** – $\hat{y}_t = a_0 + a_1 \ln t$ и др.



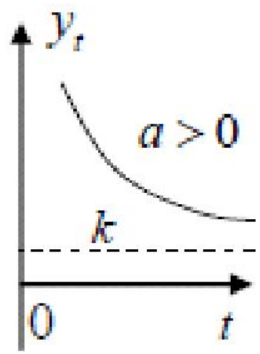
II класс – кривые, описывающие процесс, который *имеет предел роста без точки перегиба*. Такие кривые называются ***кривыми с насыщением***.

- ***кривая Джонсона*** – $\hat{y}_t = e^{a_0 + \frac{a_1}{t}}$,
- ***модифицированная экспонента*** – $\hat{y}_t = k + a_0 a_1^t$;
- ***гиперболической*** – $\hat{y}_t = a + \frac{b}{t}$;
- ***кривая Гомперца*** – $\hat{y}_t = k a_0^{a_1^t}$
- ***вторая функция Торнквиста*** – $\hat{y}_t = a_0 + \frac{t}{t + a_1}$ и др.



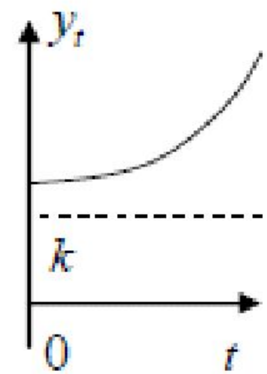
гиперболическая

$$y_t = a_0 + \frac{a_1}{t}$$



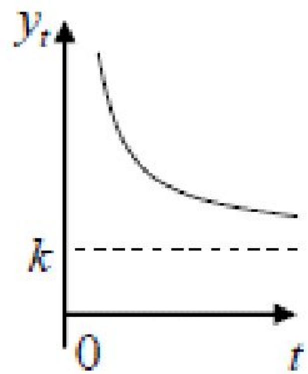
кривая Гомперца

$$y = ka^{b^t}$$



модифицированная

$$y_t = k + ab^t,$$
$$b > 1.$$



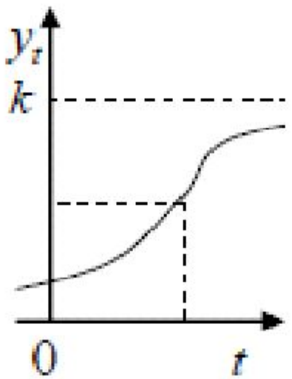
экспонента

$$y_t = k + ab^t,$$
$$0 < b < 1.$$

III класс – кривые, *имеющие предел роста и точки перегиба*.

Такие кривые называются ***S-образными***.

Описывают два последовательных лавинообразных процесса (когда прирост зависит от уже достигнутого уровня): один с ускорением развития, другой – с замедлением.



логистическая кривая

$$y(t) = \frac{k}{1 + ae^{-bt}}$$

Определение параметров кривых роста

Параметры кривых роста, как правило, оцениваются по МНК, например,

$$y_t = a_0 + a_1 t + a_2 t^2 + \ldots + a_p t^p .$$

Оценки параметров находятся в результате минимизации выражения:

$$Q = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 \to \min$$

Система нормальных уравнений:

$$\begin{cases} \sum y_t = a_0 \cdot n + a_1 \sum t + a_2 \sum t^2 + \ldots + a_p \sum t^p \\ \sum y_t \cdot t = a_0 \sum t + a_1 \sum t^2 + \ldots + a_p \sum t^{p+1} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum y_t \cdot t^{p-1} = a_0 \sum t^{p-1} + a_1 \sum t^p + \ldots + a_p \sum t^{2p-1} \\ \sum y_t \cdot t^p = a_0 \sum t^p + a_1 \sum t^{p+1} + \ldots + a_p \sum t^{2p} \end{cases}$$

Параметры прямой $y_t = a_0 + a_1 t$ **.**

Функция минимизируется, если взять частные производные по неизвестным коэффициентам

$$\begin{cases} \dfrac{\partial Q}{\partial a_0} = 2\sum\limits_{i=1}^{n}(y_i - a_0 - a_1 \cdot t_i)\cdot(-1) = 0 \\ \dfrac{\partial Q}{\partial a_1} = 2\sum\limits_{i=1}^{n}(y_i - a_0 - a_1 \cdot t_i)\cdot(-t_i) = 0 \end{cases}.$$

Система нормальных уравнений: $\begin{cases} \sum y_t = a_0 \cdot n + a_1 \sum t \\ \sum y_t \cdot t = a_0 \sum t + a_1 \sum t^2 \end{cases}$

Решение этой системы:

$$a_1 = \frac{\sum y_t \cdot t - \dfrac{\sum y_t}{n}\cdot\sum t}{\sum t^2 - \dfrac{\left(\sum t\right)^2}{n}} \qquad a_0 = \frac{\sum y_t}{n} - a_1\frac{\sum t}{n}$$

Параметры квадратичного тренда $y_t = a_0 + a_1 t + a_2 t^2$:

$$\begin{cases} \sum y_t = a_0 \cdot n + a_1 \sum t + a_2 \sum t^2 \\ \sum y_t \cdot t = a_0 \sum t + a_1 \sum t^2 + a_2 \sum t^3 \\ \sum y_t \cdot t^2 = a_0 \sum t^2 + a_1 \sum t^3 + a_2 \sum t^4 \end{cases}$$

Эта система содержит три уравнения, позволяющих найти оценки трех неизвестных коэффициентов a_0, a_1, a_2.

Величины $\sum t, \sum t^2 \ldots$ не зависят от конкретных уровней динамического ряда, они являются функциями только числа членов в ВР:

$$\sum t = \frac{n(n+1)}{2} \qquad \sum t^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum t^3 = \frac{n^2(n+1)^2}{4} \qquad \sum t^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}$$

Другой подход к упрощению расчетов – перенос начала координат в середину ВР.

для четного числа членов ряда: $t = \dots, -5; -3; -1; 1; 3; 5; \dots;$

для нечетного числа членов ряда: $t = \dots, -3; -2; -1; 0; 1; 2; 3; \dots$

Таким образом, $t_k = 0$ при k – нечетном числе.

В этом случае оценки параметров:

Прямой

$$a_1 = \frac{\sum y_t \cdot t}{\sum t^2} \qquad a_0 = \frac{\sum y_t}{n}$$

Параболы

$$a_0 = \frac{\sum y_t}{n} - \frac{\sum t^2}{n}\left[\frac{n\sum y_t \cdot t^2 - \sum t^2 \sum y_t}{n\sum t^4 - \left(\sum t^2\right)^2}\right]$$

$$a_1 = \frac{\sum y_t \cdot t}{\sum t^2} \qquad a_2 = \frac{n\sum y_t \cdot t^2 - \sum t^2 \sum y_t}{n\sum t^4 - \left(\sum t^2\right)^2}$$

Для полиномов предпочтительнее определять неизвестные коэффициенты, используя матричные обозначения.

Для параболической функции $\hat{y} = a_0 + a_1 t + a_2 t^2$.

В матричном обозначении неизвестные коэффициенты:

$$a = (T^T T)^{-1} T^T Y,$$

где $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ – вектор наблюдаемых значений временного ряда;

$T = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 \end{bmatrix}$ – матрица, характеризующая временные данные (независимые переменные).

ПРОВЕРКА АДЕКВАТНОСТИ И ТОЧНОСТИ

ВЫБРАННЫХ МОДЕЛЕЙ

Проверка адекватности выбранных моделей реальному процессу (в частности, адекватности полученной кривой роста) *строится на анализе случайной компоненты* $e_t = y_t - \hat{y}_t$.

Модель адекватна реальному явлению, если выполняются наиболее важные свойства остаточной компоненты e_t:

- равенство математического ожидания нулю: $Me_t = 0$;
- случайность остатков;
- независимость последовательных уровней ряда остатков;
- соответствие остатков нормальному закону распределения.

Проверка равенства математического ожидания нулю уровней ряда остатков

$H_0 : M(e_t) = 0$.

Для этого строится t –статистика:

$$t_{расч} = \frac{|\bar{e}|}{S} \cdot \sqrt{n} = \frac{\frac{1}{\sqrt{n}} \cdot \left| \sum_{t=1}^{n} e_t \right|}{\sqrt{\frac{1}{n-1} \sum_{t=1}^{n} (e_t - \bar{e})^2}}$$

При уровне значимости α гипотеза $H_0 : M(e_t) = 0$ отклоняется, если $t_{расч} > t_{кр}(\alpha, v = n-1)$

Здесь $t_{кр}(\alpha, v = n-1)$ – критерий Стьюдента с уровнем значимости α и $(n-1)$ степенями свободы.

Проверка случайности остатков

Для проверки случайности отклонений уровней ряда от тренда могут быть также использованы критерии:

- критерий «восходящих» и «нисходящих» серий;
- критерий пиков, или критерий поворотных точек.

Опишем второй критерий.

Значение случайной переменной считается ***поворотной точкой***, если оно одновременно больше (меньше) соседних с ним элементов:

$$e_{t-1} < e_t > e_{t+1} \text{ или } e_{t-1} > e_t < e_{t+1}\,.$$

Критерий случайности отклонения от тренда при уровне доверительной вероятности 0,95 можно представить как

$$p_{факт} > p_{расч}, \text{ где}$$

$p_{расч} = \left[(2n-1)/3 - 1{,}96\sqrt{(16n-29)/90}\right]$ – расчетное количество поворотных точек;

$p_{факт}$ – фактическое количество поворотных точек.

Проверка независимости остатков

Наиболее распространенным методом проверки наличия автокорреляции первого порядка, т.е. автокорреляции между соседними остаточными членами ряда, является критерий Дарбина–Уотсона.

Значение этого критерия определяется по формуле:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Проверка нормальности распределения остатков

При нормальном распределении показатели асимметрии (A) и эксцесса (E) равны нулю.

$$A = \frac{\frac{1}{n}\sum_{t=1}^{n} e_t^3}{\sqrt{\left(\frac{1}{n}\sum_{t=1}^{n} e_t^2\right)^3}} \qquad E = \frac{\frac{1}{n}\sum_{t=1}^{n} e_t^4}{\left(\frac{1}{n}\sum_{t=1}^{n} e_t^2\right)^2} - 3$$

$$\sigma_A = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}} \qquad \sigma_E = \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}$$

A – выборочная характеристика асимметрии;

E – выборочная характеристика эксцесса;

σ_A – среднеквадратическая ошибка выборочной характеристики асимметрии;

σ_E – среднеквадратическая ошибка выборочной характеристики эксцесса.

Если одновременно выполняются следующие неравенства:

$$\begin{cases} |A| < 1{,}5\sigma_A; \\ \left|E + \dfrac{6}{n+1}\right| < 1{,}5\sigma_E \end{cases},$$

то гипотеза о нормальном распределении случайной компоненты не отвергается.

Если выполняется хотя бы одно из неравенств

$$|A| \geq 2\sigma_A; \quad \left|E + \frac{6}{n+1}\right| \geq 2\sigma_E,$$

то гипотеза о нормальном характере распределения отвергается.

В случае попадания коэффициентов асимметрии и эксцесса в зону неопределенности между 1,5 и 2 единиц среднеквадратического отклонения

$$1{,}5\sigma_A < |A| < 2\sigma_A;$$

$$1{,}5\sigma_E < \left|E + \frac{6}{n+1}\right| < 2\sigma_E$$

используются другие критерии, в частности RS–критерий.

***RS* – критерий** – один из самых простых критериев проверки нормальности закона распределения случайной величины, он характеризует отношение размаха вариаций к стандартному отклонению $\frac{R}{S}$. Статистика *RS* рассчитывается по выборке:

$$RS_{набл} = \frac{e_{\max} - e_{\min}}{S}, \text{ где}$$

$e_t = y_t - \hat{y}_t$ – остатки, $e_{\max} = \max_t(e_t)$, $e_{\min} = \min_t(e_t)$;

$$S^2 = \frac{\sum_{t=1}^{n}(e_t - \bar{e})^2}{n-1}.$$

Если $a < RS_{набл} < b$ (***a*** – нижняя и ***b*** –верхняя границы), то гипотеза о нормальном распределении ряда остатков принимается.

Если все пункты проверки дают положительные результаты, то выбранная трендовая модель адекватна реальному ряду и ее можно использовать для построения прогнозных оценок.

Построение точечных и интервальных прогнозов

Доверительный интервал определяется в виде:

$$\hat{y}_{n+\tau} \pm t_{\alpha,v} S_{np} \text{ , где}$$

n – длина временного ряда;

τ – период упреждения;

$\hat{y}_{n+\tau}$ – точечный прогноз на момент $n+\tau$;

$t_{\alpha,v}$ – значение критерия Стьюдента (α – уровень значимости и $v=n-k-1$ –степень свободы, k – количество параметров при независимых переменных);

S_{np} – средняя квадратическая ошибка прогноза.

Стандартная ошибка прогноза зависит не только от числа наблюдений n, но и от периода упреждения τ и может быть определена в матричной форме

$$S_{np} = S \cdot \sqrt{1 + T_\tau^T (T^T T)^{-1} T_\tau} \ ,$$

где S – стандартная ошибка уравнения; T_τ – вектор-столбец времени, по которому производится экстраполяция:

$\hat{y}_t = a_0 + a_1 t$: $$T_\tau = \begin{pmatrix} 1 \\ n+\tau \end{pmatrix}$$

$\hat{y}_t = a_0 + a_1 t + a_2 t^2$: $$T_\tau = \begin{pmatrix} 1 \\ n+\tau \\ (n+\tau)^2 \end{pmatrix}$$

$$S = \sqrt{\frac{Q}{n-k-1}} \text{ и } Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \text{, где}$$

k – количество параметров при независимых переменных.