Семинар 4. Временные ряды. Автокорреляционная функция

4.1. Пример временного ряда

Рассмотрим пример: серия измерений давления газа на выходе из абсорбера на УКПГ. На первый взгляд, давление P линейно зависит от времени t,

$$P_t = \beta_0 + \beta_1 t + \varepsilon_t,$$

где ε_t – случайная ошибка измерения, белый шум.

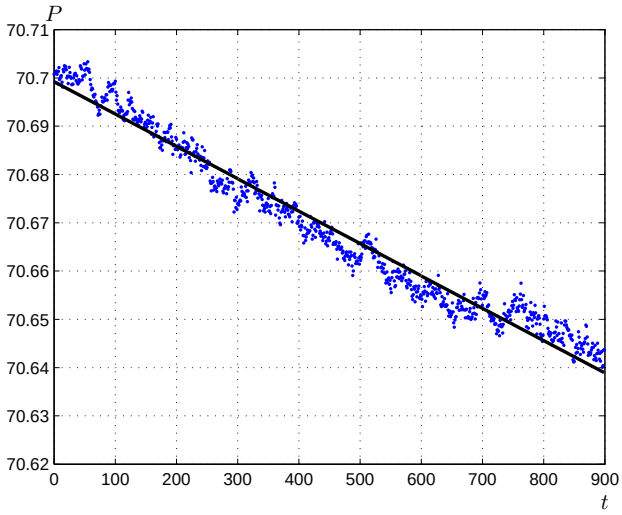

Рис. 4.1. Линейная модель зависимости давления от времени

Однако, остатки этой модели далеки от белого шума.

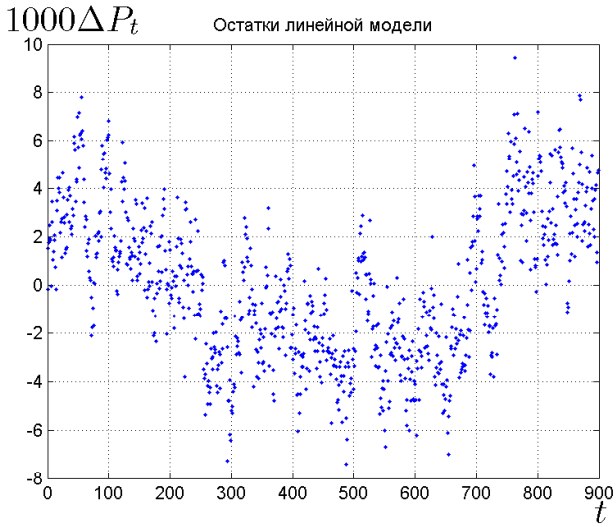

Рис. 4.2. Остатки линейной модели

Чтобы понять, что это значит, нужно объяснить, что такое белый шум, остатки и автокорреляционная функция. Поэтому перейдем к теории.

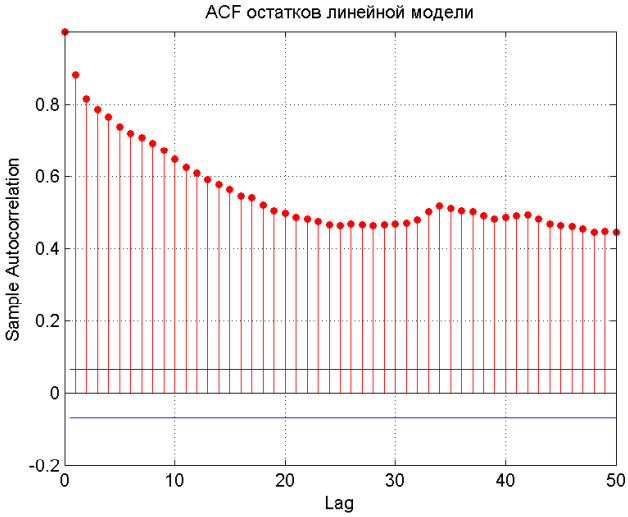

Рис. 4.3. Автокорреляционная функция остатков линейной модели

4.2. Автокорреляционная функция (ACF)

В рассмотренном примере серия давлений, измеренных через равные промежутки времени, – это *временной ряд*. В общем случае временной ряд – это бесконечная в обе стороны последовательность $..., y_{t-2}, y_{t-1}, y_t, y_{t+1}, y_{t+2}, ...$ с целыми номерами t.

На каждое случайное явление может быть два взгляда: *до* измерения и *после* него. До измерения явление характеризуется случайными величинами, их распределениями, математическими ожиданиями и проч., т.е. вероятностной моделью. После измерения случайные величины превращаются в числа, с помощью которых мы проверяем на эти модели на соответствие действительности.

Таким образом, до измерений временной ряд – это бесконечная в обе стороны последовательность случайных величин, которые могут как-то зависеть друг от друга. После измерений – это конечная выборка чисел с номерами $1, 2, ..., n$. Анализ временного ряда преследует следующие цели:

- подобрать модель формирования ряда;
- спрогнозировать будущее поведения ряда;
- учесть погрешность прогноза (построить доверительный интервал);
- смоделировать аналогичные ряды для тестирования управляющих устройств.

Когда мы можем считать, что временной ряд задан? Как можно их сравнивать друг с другом? Когда считать два ряда одинаковыми?

Если ряд рассматривается после измерений, то тут вопросов нет. Если все числа с одинаковыми номерами совпадают, то и ряды равны. А до измерений?

До измерений надо сравнивать вероятностные распределения. Если у нас есть 2 отдельные случайные величины, то они считаются равными при совпадении их функций распределения,

$$\mathbf{P}\{X_1 \le x\} = \mathbf{P}\{X_2 \le x\}$$

для всех допустимых значений x.

Когда рассматривается временной ряд, то этого не достаточно, т.к. могут быть взаимосвязи между членами ряда. С точки зрения математики временной ряд – это случайный процесс с дискретным временем. Из теории случайных процессов известно, что процесс задается совместными распределениями всех сочетаний членов ряда. Т.е. нужно знать не только распределения каждого

члена ряда, но и всех пар, всех троек, всех четверок и т.д. Это совершенно необозримо и невозможно реализовать практически. Поэтому теория случайных процессов занята исследованием таких классов случайных процессов, которые можно задать как-то попроще.

Один из таких классов – гауссовские временные ряды. В этом случае совместное распределение нескольких членов ряда имеет многомерное нормальное распределение,

$$f(x_1,\dots x_r)=\frac{1}{\sqrt{(2\pi)^r \det C}}\exp\{-\frac{1}{2}(\vec{x}-\vec{\mu})'C^{-1}(\vec{x}-\vec{\mu})\},$$

где $\vec{\mu}$ – вектор математических ожиданий, C – ковариационная матрица.

Таким образом, для того, чтобы определить ряд, нам достаточно знать мат. ожидания членов ряда и их ковариации

$$C_{ts}=\mathrm{cov}(y_t,y_s)=\mathbf{M}(y_t-\mu_t)(y_s-\mu_s).$$

Это уже большое облегчение, но надо еще сужать круг исследуемых объектов. Рассмотрим стационарные временные ряды.

Временной ряд y_t называется *стационарным*, если его совместные распределения не зависят от времени t, а зависят только от сдвига – расстояния между членами ряда.

Гауссовский ряд y_t будет стационарным, если его математическое ожидание $\mathbf{M}y_t$, дисперсия $\mathbf{D}y_t$ и автоковариация $\mathrm{cov}(y_t,y_{t-k})$ не зависят от времени (номера наблюдения) t.

Если ряд не гауссовский, но для него выполняются эти условия, то он называется *стационарным в широком смысле*, а вся теория гауссовских рядов применяется к нему как приближенная.

Содержательное отличие временных рядов друг от друга состоит именно во взаимной зависимости членов ряда, а не в значении параметров сдвига ($\mu=\mathbf{M}y_t=\mathrm{const}$) и масштаба ($\sigma^2=\mathbf{D}y_t=\mathrm{const}$). Поэтому стационарные ряды всегда считаются центрированными ($\mu=0$). Логично было бы и нормировать ряд ($\sigma^2=1$), но это почему-то делать не принято. Нормируют автоковариацию – получается автокорреляция.

Поскольку автоковариация зависит только от сдвига, то и автокорреляция – это функция только сдвига k,

$$\rho(k)=\frac{\mathrm{cov}(y_t,y_{t-k})}{\sigma^2},k\geq 0.$$

Это и есть та самая автокорреляционная функция (ACF), которую мы использовали в примере. Она полностью (с точностью до сдвига и масштаба) определяет все свойства стационарного гауссовского ряда.

Самый простой вид зависимости – ее полное отсутствие, т.е. $\rho(k)\equiv 0$ при $k>0$ ($\rho(0)=1$ всегда). Такой ряд называется *белым шумом* и обычно обозначается ε_t.

Из белого шума можно получить любой другой гауссовский стационарный ряд с помощью *линейного фильтра*,

$$y_t=\sum_{i=0}^{\infty}\alpha_i\varepsilon_{t-i}.$$

Чтобы такой ряд был стационарным, необходимо и достаточно, чтобы его дисперсия была конечная,

$$\mathbf{D}y_t=\sum_{i=0}^{\infty}\alpha_i^2\mathbf{D}\varepsilon_{t-i}=\sigma^2\sum_{i=0}^{\infty}\alpha_i^2,$$

т.е. ряд из квадратов коэффициентов должен сходиться.

Найдем автоковариации отфильтрованного шума,

$$\mathrm{cov}(y_t,y_{t-k})=\mathrm{cov}\left(\sum_{i=0}^{\infty}\alpha_i\varepsilon_{t-i},\sum_{i=0}^{\infty}\alpha_i\varepsilon_{t-k-i}\right)=\mathrm{cov}\left(\sum_{i=0}^{\infty}\alpha_i\varepsilon_{t-i},\sum_{i=k}^{\infty}\alpha_{i-k}\varepsilon_{t-i}\right)=\sigma^2\sum_{i=k}^{\infty}\alpha_i\alpha_{i-k},$$

поскольку ковариации ε_{t-i} с разными номерами равны нулю. Сдвинув индекс суммирования к нулю и разделив на σ^2, получим

$$\rho(k)=\sum_{i=0}^{\infty}\alpha_i\alpha_{i+k}.$$

Поскольку ряд $\sum\limits_{i=0}^{\infty}\alpha_i^2$ сходится, то $\alpha_i \to 0$ при $i \to \infty$. Следовательно, с ростом k автокорреляционная функция стационарного ряда должна стремиться к нулю, $\rho(k) \to 0$.

Коэффициенты α_k имеют смысл корреляций ряда с белым шумом,

$$\operatorname{cov}(y_t, \varepsilon_{t-k}) = \sum_{i=0}^{\infty}\alpha_i \operatorname{cov}(\varepsilon_{t-i}, \varepsilon_{t-k}) = \alpha_k \operatorname{cov}(\varepsilon_{t-k}, \varepsilon_{t-k}) = \alpha_k\sigma^2.$$

Отсюда

$$\alpha_k = \frac{\operatorname{cov}(y_t, \varepsilon_{t-k})}{\sigma^2}.$$

Все это был взгляд "до измерений". После измерений автокорреляционной функции уже нет, есть выборка числовых значений. По ней можно вычислить *выборочные автоковариации*, а затем получить выборочную автокорреляционную функцию,

$$\begin{aligned}\widehat{\rho}(k) &= \frac{1}{(n-k)\widehat{\sigma}^2}\sum_{i=k+1}^{n} y_i\, y_{i-k}\\ \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} y_i^2.\end{aligned}$$

Если она обладает теми же свойствами, что и ACF какой-либо модели, то мы будем считать, что эта модель хорошо описывает механизм формирования выборки.

Однако из-за конечности выборки эти показатели вычисляются с некоторой случайной погрешностью, для учета которой применяются методы математической статистики. Во всех моделях предполагается, что ошибки – это белый шум, поэтому основная задача – проверка гипотезы о равенстве автокорреляционной функции нулю.

Если истинная автокорреляция равна нулю, $\rho(k) = 0$, то выборочная автокорреляция имеет приближенно нормальное распределение с нулевым мат. ожиданием и дисперсией $1/n$,

$$\widehat{\rho}(k) \in N\left(0, \frac{1}{\sqrt{n}}\right),$$

т.е. выборочные автокорреляции должны лежать в диапазоне $\left[-\dfrac{z_{1-\alpha/2}}{\sqrt{n}},\ \dfrac{z_{1-\alpha/2}}{\sqrt{n}}\right]$ с вероятностью $1-\alpha$. Обычно берут $\alpha = 0.05$, поэтому квантиль стандартного нормального распределения $z_{0.975} \approx 2$.

Этот способ проверки несколько грубоват. Более точен Q-критерий Льюнга-Бокса,

$$\begin{aligned}H_0 &: \quad \rho_i = 0, \quad i = \overline{1,k},\\ Q &= n(n+2)\sum_{i=1}^{k}\frac{\widehat{\rho}_i^2}{n-i} \to \chi_k^2 \quad \text{при } n \to \infty.\end{aligned}$$

Методику проверки гипотез можно найти в любом учебнике по статистике, не будем на ней сейчас подробно останавливаться.

Подведем небольшой итог.

1. Если выборочная ACF исходного ряда не стремится к нулю (с точностью до $\dfrac{2}{\sqrt{n}}$), то ряд нестационарный. Его нужно каким-либо способом свести к стационарному.

2. При выборе модели следует смотреть на форму ACF стационарного ряда и сравнивать ее с ACF разных моделей. На что больше похоже – то и пробовать.

3. Для этого нужно знать вид ACF для широкого круга моделей. Чем больше моделей мы знаем, тем точнее сможем подобрать.

4. Если в модели есть ошибки в виде белого шума (а они есть практически везде), то остатки тоже должны обладать свойствами белого шума, т.е. их ACF должна быть нулевой с точностью до $\dfrac{2}{\sqrt{n}}$. Уточнить можно с помощью критерия Льюнга-Бокса.

4.3. Авторегрессионные модели

Вернемся к примеру. Можно предложить другую модель зависимости – случайное блуждание,

$$P_t = P_{t-1} + \varepsilon_t,$$

Тогда ряд разностей $\Delta P_t = P_t - P_{t-1}$ должен оказаться белым шумом.

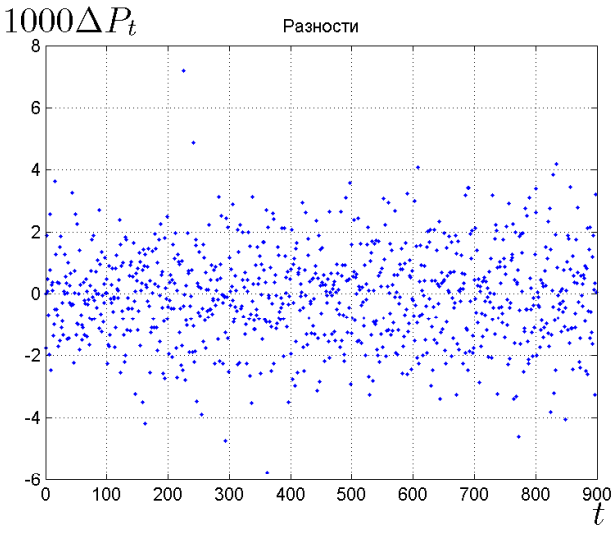

Рис. 4.4. Разности первого порядка

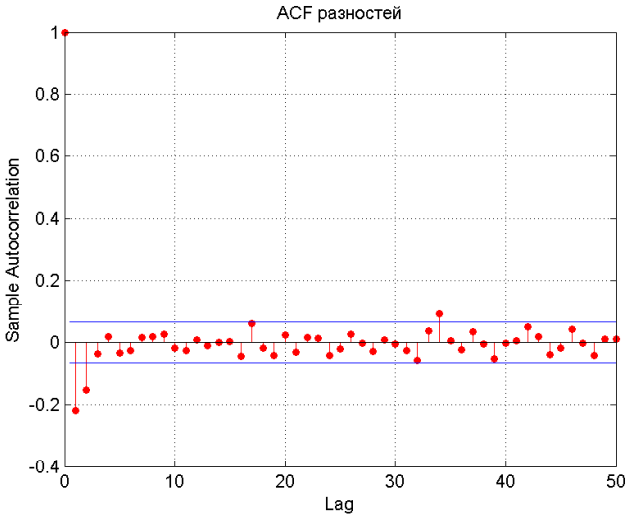

Рис. 4.5. Автокорреляционная функция разностей

Это больше похоже на белый шум, но есть некоторые от него отклонения. Проверить это поможет Q-критерий Льюнга-Бокса. Вычислим значимость – вероятность того, что случайная величина Q ("до измерений") окажется больше ее реализации $\widehat{Q}$, вычисленной по выборке ("после измерений"),

$$p = \mathbf{P}\{Q > \widehat{Q}\} = 5.3 \cdot 10^{-11}.$$

Это значительно меньше общепринятых уровней значимости (0.1; 0.05; 0.01), поэтому гипотеза об отсутствии автокорреляции должна быть отвергнута. Следовательно, нужно еще усложнить модель. Возьмем модель авторегрессии 7-го порядка (с некоторыми исключенными слагаемыми) для исходных давлений,

$$P_t = a_1\, P_{t-1} + a_3\, P_{t-3} + a_7\, P_{t-7} + \varepsilon_t,$$

или модель авторегрессии 5-го порядка (без всяких исключений) для разностей,

$$\Delta P_t = \sum_{i=1}^{5} a_i \, \Delta P_{t-i} + \varepsilon_t.$$

В обоих случаях гипотеза об отсутствии автокорреляции остатков для первых 10-ти лагов не отвергается ($p = 0.34$ для первой модели и $p = 0.61$ для второй). Приведем, например, ACF остатков первой модели.

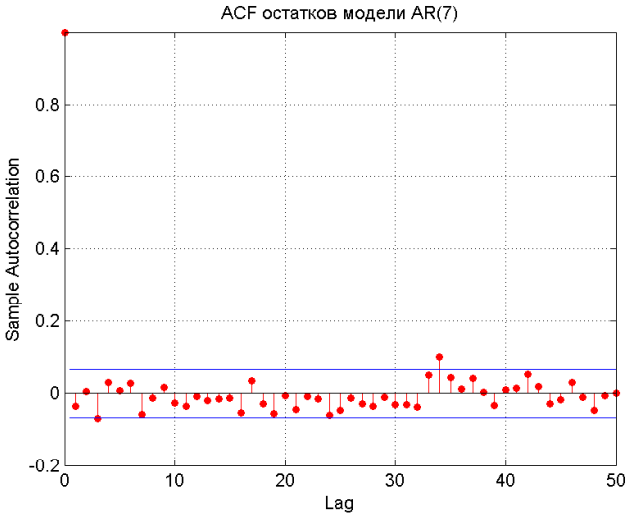

Рис. 4.6. Автокорреляционная функция остатков AR(7)

4.4. Задачи

По серии измерений давления на выходе абсорбера постройте рассмотренные выше модели:

- линейную;
- разности первого порядка;
- авторегрессию для исходных давлений; подберите минимальный набор авторегрессионных факторов, обеспечивающих отсутствие автокорреляции остатков;
- авторегрессию для разностей давлений; подберите минимальный набор авторегрессионных факторов, обеспечивающих отсутствие автокорреляции остатков.

4.5. Вопросы для самоконтроля

1. Что такое белый шум? Как его отличить от других временных рядов?
2. Почему исследуют именно ACF, а не другие вероятностные характеристики временного ряда?
3. Может ли выборочная ACF в точности равняться нулю?
4. Как по ACF отличить стационарный ряд от нестационарного? Почему?