

Лектор – *Сенько Олег Валентинович*

Курс «ПРИКЛАДНОЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ. ЧАСТЬ II»
ПС

Эффективным непараметрическим инструментом статистической верификации является перестановочный тест. Целью перестановочного теста является проверка существования зависимости между целевой переменной Y от и набором переменных X_1, \dots, X_n в рамках некоторой гипотезы о соответствующей математической модели. Проверка существования зависимости сводится к попытке опровержения нулевой гипотезы о независимости переменной Y от переменных X_1, \dots, X_n с использованием статистики, характеризующей качество аппроксимации зависимости с использованием предложенной модели. Например, при гипотезе о существовании прямой линейной связи между Y и одной единственной переменной X статистикой может являться коэффициент корреляции Пирсона. При гипотезе о существовании линейной связи между Y и X_1, \dots, X_n статистикой критерия может служить величина R^2 .

Предположим, что у нас имеется выборка $\tilde{S} = \{(y_1, \mathbf{x}_1), \dots, (y_m, \mathbf{x}_m)\}$. Для использования перестановочного теста необходимо:

- Сделать предположение о характере зависимости и предложить соответствующую этому предположению статистику T .
- Рассчитать значение статистики T для исходной выборки \tilde{S}
- Выбрать количество перестановок N исходя из имеющихся вычислительных ресурсов.
- Приравнять 0 целочисленный показатель N_t , подсчитывающий достижение или превышение статистики критерия при справедливости нулевой гипотезы H_0 (см. далее)
- А) С использованием генератора случайных чисел получить случайную перестановку f чисел из набора $\{1, \dots, m\}$

- В) По перестановке f построить случайную выборку $\tilde{S}_r = \{(y_{f(1)}, \mathbf{x}_1), \dots, (y_{f(m)}, \mathbf{x}_m)\}$
- С) Рассчитать значение статистики T для случайной выборки \tilde{S}_r
- D) При выполнении неравенства $T(\tilde{S}_r) \geq T(\tilde{S})$ $N_t = N_t + 1$. В противном случае N_t не меняется.
- Независимо повторить пункты A)-D) N раз.
- В качестве p -значения использовать отношение $\frac{N_t}{N}$

Предположим, что \tilde{S}_r^j - выборка, полученная из исходной выборки \tilde{S} с использованием случайной перестановки чисел из $\{1, \dots, m\}$ с порядковым номером j . Приведённую выше процедуру подсчёта p -значения можно выразить формулой

$$p = \frac{\sum_{j=1}^N I[T(\tilde{S}_r^j) \geq T(\tilde{S})]}{N}, \quad (1)$$

где $I[b] = 1$ при $b = true$ и $I[b] = 0$ в противном случае.

В классической статистике p –значение определяется как $P(T \geq |T(S)|H_0)$ При этом в H_0 наряду с предположением о характеристиках распределений, из которых генерируются данные, входит также предположение о независимости наблюдений. При использовании перестановочного теста в нулевую гипотезу входят следующие предположения о процессе, генерирующем данные:

- Все генерируемые выборки имеют один и тот же размер m .
- Вектора переменных X_1, \dots, X_n считаются детерминированными, то есть n вектора X -описаний для всех генерируемых выборок одинаковы и равны $\mathbf{x}_1, \dots, \mathbf{x}_m$
- значения Y генерируются случайно и независимо из маргинальных эмпирического распределения \hat{E}_y . Такая генерация совпадает с выборкой с возвращением.

Нулевая гипотеза H_0 заключается в предположении о генерации данных описанным выше процессом.

- Процесс отбирает выборки значений переменной Y , которые совпадают по размеру с выборкой \tilde{S} и распределению на ней значений Y

Theorem

Предположим, что $N_1^p[T(\tilde{S})]$ число таких отличающихся друг от друга перестановок y -частей описаний объектов из \tilde{S} при фиксированных x -частях, при которых для результирующей выборки \tilde{S}' справедливо неравенство $T(\tilde{S}') \geq T(\tilde{S})$. Тогда справедливо равенство

$$P[T(\tilde{S}') \geq T(\tilde{S}) | H_0] = \frac{N_1^p[T(\tilde{S})]}{m!} \quad (2)$$

Доказательство Обозначим через $W^p(\tilde{S})$ множество выборок, генерируемых описанным выше процессом. Разобьём $W^p(\tilde{S})$ на группы неразличимых выборок $\tilde{G}_1, \dots, \tilde{G}_{N_d}$. Выборки \tilde{S}' и \tilde{S}'' считаются тождественными, если значения Y на объектах с одинаковым номером равны. Доказательство опирается на 3 утверждения.

- 1) Вследствие неразличимости величина статистики T на всех выборках внутри каждой из групп постоянна.
- 2) Число перестановок позиций y —частей описаний выборки при фиксированных x —частях, с помощью которых могут быть получены все выборки из группы является одинаковым для всех групп.
- 3) Вероятности появления каждой из групп одинаковы.

Доказательство пункта 1) очевидно.

Докажем пункт 2). Предположим, что Y в \tilde{S} принимает значения $k \leq t$ значений. Пусть l_1, \dots, l_k являются количествами объектов в \tilde{S} , на которых Y принимает значения $\check{y}_1, \dots, \check{y}_k$ соответственно. Любая перестановка, переводящая \tilde{S} в произвольную перестановку из группы G_u является произведением некоторой перестановки π_{0u} , осуществляющей переход в одну из выборок из G_u и перестановки из $\tilde{\pi}_u$, где $\tilde{\pi}_u$ является множеством перестановок, сохраняющих принадлежность G_u . Число перестановок в $\tilde{\pi}_u$ не зависит от номера группы u и равна $T_g = \prod_{i=1}^k l_i!$.

При случайной и независимой генерации Y из маргинальных эмпирического распределения \hat{E}_y вероятности всех групп также равны между собой $\prod_{i=1}^k P^{l_i}(\check{y})$. Дополнительный отбор сохраняет равновероятность. Вследствие этого $P(G_u) = \frac{1}{N_d}$ для всех u .

$$P[T(\tilde{S}') \geq T(\tilde{S})|H_0] = \frac{\sum_{u \in U[T(\tilde{S})]} P(G_u)}{\sum_{u=1}^{N_d} P(G_u)}, \quad (3)$$

где $U[T(\tilde{S})]$ -множество номеров групп, для выборок которых выполняется условие $T(\tilde{S}') \geq T(\tilde{S})$. Домножим числитель и знаменатель левой части равенства (3) на $T_g N_d$. В результате и получаем равенство (2).

На практике вместо подсчёта всевозможных перестановок используется их подмножество, получаемое с помощью случайной генерации перестановок. Использование нескольких тысяч перестановок позволяет достаточно точно оценивать p -значения.

- **Преимущества.** Безусловно преимуществом перестановочного теста является отсутствие требований к типу вероятностных распределений, а также отсутствие требований к размеру выборок. Преимуществом также является гибкость. Для использования необходимо только предположить вид зависимости и подобрать характеризующую данную зависимость статистику. Не требуется аналитически восстанавливать распределение статистики.
- **Недостатки** Основным недостатком перестановочного теста является привязанность его к конкретным данным. Тест может приводить к ошибочным заключениям при сильном, но статистически возможным отклонении от генеральной совокупности. Другим недостатком является необходимость больших объемов вычислений.

Целью многих исследований в различных областях науки является выяснение, какие показатели из некоторого заранее заданного набора X_1, \dots, X_n связаны с целевой переменной Y .

Стандартной процедурой установления связи переменной X_j с Y является проверка нулевой гипотезы H_0^j об отсутствии связи, которая сводится к сравнению рассчитанного p -значения с некоторым заранее заданным уровнем значимости α : H_0^j отвергается, если выполняется неравенство $p \leq \alpha$.

При применении указанной процедуры к одной единственной переменной X_j вероятность ошибки первого рода, то есть вероятность ошибочного опровержения нулевой гипотезы равна α .

Нулевые гипотезы об отдельных эффектах, связанных с конкретными переменными далее будем называть индивидуальными. Допустим, что индивидуальная нулевая гипотеза отвергнута для хотя бы одной переменных из X_1, \dots, X_n . В этом случае уверенность вывода о действительном существовании найденных связей с Y связана с вероятностью ошибочного отвержения по крайней мере одной из множества на самом деле верных индивидуальных нулевых гипотез. Такую вероятность принято называть family wise error rate (FWER). Обозначим через V общее число ошибочно отвергнутых нулевых гипотез. Тогда $FWER = P(V \geq 1)$. Очевидно, что FWER зависит от уровня значимости, на котором отвергаются индивидуальными гипотезы.

Наряду с FWER для оценки множественного тестирования могут быть использованы также другие меры. Пусть R - общее число отвергнутых индивидуальных нулевых гипотез.

- PCER (per-comparison error rate) - математическое ожидание доли ошибок первого рода $\frac{\mathbb{E}(V)}{n}$
- PFER(per-family error rate) - математическое ожидание ошибок первого рода $\mathbb{E}(V)$
- FDR(false discovery rate) - математическое ожидание доли ошибок первого рода среди отвергнутых индивидуальных нулевых гипотез. Поскольку R может принимать нулевые значения, то FDR определяется как $\mathbb{E}(\frac{V}{R} | R > 0)P(R > 0)$

Аналогично тому, как для оценивания значимости отдельных связей проверяется неравенство $p \leq \alpha$, для оценки значимости с учётом эффекта множественного тестирования может проверяться неравенство $FWER \leq \alpha$. Попробуем оценить вероятность случайного отвержения по крайней мере одной из нулевых гипотез H_0^1, \dots, H_0^n на уровне не хуже β .

Обозначим через $\Omega_j^{r\beta}$ -множество выборок данных, генерируемых при справедливости нулевой гипотезы H_0^j , при которых эта гипотеза отвергается на уровне β .

FWER, то есть вероятность, что хотя бы одна одна из гипотез H_0^j будет ошибочно отвергнута на уровне β очевидно равна $P(\cup_{i=1}^n \Omega_j^{r\beta})$

Проблема множественного тестирования. Коррекция по Бонферрони

Для оценки сверху вероятности $P(\cup_{i=1}^n \Omega_j^{r\beta})$ может быть использовано неравенство Буля

$$P(\cup_{i=1}^n \Omega_j^{r\beta}) \leq \sum_{i=1}^n P(\Omega_j^{r\beta})$$

Однако $P(\Omega_i^{r\beta}) = \beta$ по определению множества $\Omega_i^{r\beta}$. Поэтому

$$P(\cup_{i=1}^n \Omega_j^{r\beta}) \leq n\beta$$

Откуда следует, что для выполнения требования $FWER \leq \alpha$ достаточно, чтобы выполнялось неравенство $n\beta \leq \alpha$ или $\beta \leq \frac{\alpha}{n}$. Определяемое последним неравенством требование к уровню значимости, на котором отвергаются индивидуальные нулевые гипотезы, носит название **поправки Бонферрони**.

Проблема множественного тестирования. Коррекция по Бонферрони

Требование выполнения неравенства $\beta \leq \frac{\alpha}{n}$, эквивалентно использованию вместо исходных p –значений скорректированных p –значений. Скорректированное p –значение для нулевой гипотезы H_0^j вычисляется через исходное p –значение как $\tilde{p}_j = \min(np_j, 1)$.

Скорректированное p –значение является минимальной ошибкой первого рода, при которой нулевая гипотеза отвергается с учётом поправки на множественное тестирование.

Видно, метод Бонферрони выдвигает весьма жёсткие требования к уровню значимости, на котором отвергается индивидуальная нулевая гипотеза. Например, для положительного заключения о существовании связи между переменной X_j и целевой переменной Y на уровне 0.01 необходимо, чтобы соответствующая индивидуальная нулевая гипотеза была отвергнута на уровне $\frac{0.01}{n}$.

Метод Шидака. В подходе Шидака предполагается, что p -значения p_1, \dots, p_n , соответствующие индивидуальными нулевым гипотезам H_0^1, \dots, H_0^n , являются взаимонезависимыми случайными величинами, подчиняющимися равномерному распределению на отрезке $[0, 1]$. Скорректированное p -значение в методе Шидака вычисляется по формуле

$$\tilde{p}_j = 1 - (1 - p_j)^n$$

Оценки по методу Шидака являются корректными только при выполнении неравенства Шидака для совместного распределения статистик, используемых оценивании индивидуальные нулевые гипотезы

$$Pr(|T_1| \leq c_1, \dots, |T_n| < c_n) \geq \prod_{j=1}^n Pr(|T_j| \leq c_j).$$

Неравенство Шидака справедливо в частности для многомерных нормальных распределений.

Проблема множественного тестирования. Одношаговые процедуры

Недостатком методов Бонферрони и Шидака является учёт при вычислении скорректированной значимости отвержения нулевой гипотезы H_0^j только лишь исходного нескорректированного p –значения, а также общего числа исходных нулевых гипотез. При этом не учитываются исходные p –значения для остальных индивидуальных нулевых гипотез помимо H_0^j . Подобные процедуры принято называть одношаговыми (single step). Одной из возможных способов задания пошаговых процедур является подход Вестфолла и Янга (Westfall and Young) в котором рассчитанные p –значения рассматриваются как реализации случайных величин P_1, \dots, P_n . При этом скорректированное значение определяется по формуле

$$\tilde{p}_i = P(\min_{1 \leq l \leq n} P_l \leq p_i | H_u), \quad (4)$$

Проблема множественного тестирования. Одношаговые процедуры

где H_u является пересечением всех индивидуальных нулевых гипотез $H_u = \cap_{j=1}^n H_0^j$. В случае, когда для всех индивидуальных нулевых гипотез распределения статистик критерия близки скорректированное значение может приближённо вычисляться по формуле:

$$\tilde{p}_i = P(\max_{1 \leq l \leq n} T_l \geq t_i | H_u). \quad (5)$$

Предполагается, что рассчитанные значения статистик критерия для каждой из нулевых гипотез t_1, \dots, t_n являются реализация случайных функций T_1, \dots, T_n соответственно.

Для вычисления скорректированных значений согласно формулам (4,5) может быть использован также перестановочный тест.

Предположим, что у нас имеется обучающая выборка $\tilde{S} = \{(y_1, \mathbf{x}_1) \dots, (y_m, \mathbf{x}_m)\}$. Требуется проверить n нулевых гипотез H_0^1, \dots, H_0^n о независимости целевой переменной Y от каждой из переменных X_1, \dots, X_n с помощью некоторой статистики T . Сгенерируем множество \tilde{f}_N из N случайных независимых перестановок чисел из $\{1, \dots, m\}$: $\tilde{f}_N = \{f_j | j = 1, \dots, N\}$. Для выборки $\tilde{S}_r^j = \{(y_{f_j(1)}, \mathbf{x}_1), \dots, (y_{f_j(m)}, \mathbf{x}_m)\}$ вычислим статистику

$$T_{max}^j = \max_{i \in \{1, \dots, n\}} T_i(\tilde{S}_r^j).$$

Скорректированное согласно формуле (5) p -значение может быть оценено с помощью модифицированного варианта формулы 1:

$$\tilde{p}_i = \frac{\sum_{j=1}^N I[T_{max}^j \geq T_i(\tilde{S})]}{N}.$$

Пусть $p_1(\tilde{S}), \dots, p_n(\tilde{S})$ являются p -значениями, рассчитанными по выборке \tilde{S} с помощью описанной ранее процедуры, соответствующей формуле (1). Скорректированное согласно формуле (4) p -значение может быть оценено как

$$\tilde{p}_i = \frac{\sum_{j=1}^N I[\hat{P}_{min}^j \leq p_i(\tilde{S})]}{N},$$

где $\hat{P}_{min}^j = \min_{i \in \{1, \dots, n\}} p_i(\tilde{S}_r^j)$

Для получения более точных оценок скорректированной значимости эффекта, связанного с отвержением индивидуальной нулевой гипотезы H_0^j , необходимо учитывать исходные p -значения для всех индивидуальных нулевых гипотез. Проводить такой учёт позволяют процедуры пошагового спуска (step down). Одной из таких процедур является метод **метод Бонферрони-Холма**.

Предположим, что для нулевых гипотез H_0^1, \dots, H_0^n рассчитаны p -значения p_1, \dots, p_n . Пусть $p^{(1)}, \dots, p^{(n)}$ - ряд упорядоченных по возрастанию p -значений и $H_0^{(1)}, \dots, H_0^{(n)}$ - ряд соответствующих нулевых гипотез.

Зафиксируем уровень значимости α .

Предположим, что h - минимальный индекс, удовлетворяющий условию:

$$p^{(h)} > \frac{\alpha}{n+1-h}. \quad (6)$$

Тогда нулевые гипотезы $H_0^{(1)}, \dots, H_0^{(h-1)}$ отвергаются, а нулевые гипотезы H_0^h, \dots, H_0^n принимаются. В случае, когда $h = 1$ принимаются все нулевые гипотезы. В случае, когда не существует $h \in 1, \dots, m$, при котором выполняется неравенство (6), то отвергаются все нулевые гипотезы H_0^1, \dots, H_0^n .

Для оценки скорректированного уровня значимости с учётом всей совокупности исходных p —значений для индивидуальных нулевых гипотез также может быть использован индекс FDR. Предположим, что используется процедура, при которой индивидуальные нулевые гипотезы отвергаются, если соответствующее исходное p —значение не превышает β . Оценим предполагаемое число ошибок первого рода при условии справедливости всех нулевых гипотез H_0^1, \dots, H_0^n . Если предполагать независимость соответствующих статистик, то в качестве оценки может быть использовано произведение βn .

Предположим, что настоящее число отвергнутых нулевых гипотез составило m' .

Опровержение этих гипотез нельзя считать случайным, если βn многократно меньше n' . Предположим, что

$$\frac{\beta n}{n'} \leq \delta \quad (7)$$

Неравенство (7) означает, что ожидаемая доля ошибок первого рода среди отвергнутых гипотез не превышает δ . Естественно оценивать на уровне δ скорректированную значимость индивидуальной гипотезы, которая исходно отвергается на уровне β . Величина $\frac{\beta n}{n'}$ фактически является оценкой FDR.

По **временным рядом** понимается совокупность наблюдений некоторой величины X в различные моменты времени из некоторого интервала $[0, T]$. При этом чаще всего предполагается, что время наблюдения t принимает целочисленные значения из отрезка $[0, T]$. Предполагается, что каждый рассматриваемый временной ряд является реализацией **дискретного случайного или стохастического процесса**.

Каждому целочисленному моменту времени $t \in [0, T]$ ставится в соответствие случайная величина X_t . Дискретным стохастическим процессом понимается совокупность всех таких случайных величин. Временной ряд называется реализацией дискретного стохастического процесса, если значение ряда в момент времени t является реализацией случайной величины X_t .

Часто имеет смысл трактовать последовательность $\{X_t, t \in [0, T]\}$ как подпоследовательность бесконечной последовательности $\{X_t, t = \dots, -2, -1, 0, 1, 2, \dots\}$.

Очевидно, что стохастический процесс полностью характеризуется совокупностью плотностей совместных распределений случайных величин вида $f(x_{t_1}, \dots, x_{t_n})$, где t_1, \dots, t_n — некоторый набор моментов наблюдений.

Стохастический процесс называется **строго стационарным** или **стационарным в узком смысле**, если для произвольного набора моментов наблюдения t_1, \dots, t_n и для произвольного целочисленного Δ справедливо равенство

$$f(x_{t_1}, \dots, x_{t_n}) = f(x_{t_1+\Delta}, \dots, x_{t_n+\Delta})$$

Из строгой стационарности следует

- Независимость от t математического ожидания $\mathbb{E}X_t$, то есть $\forall t \in \mathbb{Z} \mathbb{E}X_t$ равно некоторой одной и той же величине μ

- Независимость от t дисперсии $\mathbb{E}(X_t - \mu)^2$, то есть $\forall t \in \mathbb{Z}$ $\mathbb{E}(X_t - \mu)^2$ равно некоторой одной и той же величине σ^2
- Автоковариационная функция (значение ковариации между случайными величинами, соответствующими двум разными моментами времени t_1 и t_2) зависит только от разности $t_1 - t_2$

Последний вывод следует из выполнения при произвольном целочисленном Δ равенства

$$\begin{aligned} \text{Cov}(X_{t_1}, X_{t_2}) &= \int \int (x_{t_1} - \mu)(x_{t_2} - \mu) f(x_{t_1}, x_{t_2}) dx_{t_1} dx_{t_2} = \\ &= \int \int (x_{t_1} - \mu)(x_{t_2} - \mu) f(x_{t_1+\Delta}, x_{t_2+\Delta}) dx_{t_1} dx_{t_2} \end{aligned}$$

Под **слабой стационарностью** или **стационарностью процесса в широком смысле** понимается процесс, для которого математическое ожидание и дисперсия не зависят от времени, а автоковариационная функция зависит только от разности моментов времени $t_1 - t_2$. Слабая стационарность также называется иногда также стационарностью в ковариациях.

Под **белым шумом** понимается случайный процесс, в котором $X_t = \epsilon_t$, где $\forall t \mathbb{E}\epsilon_t = 0$, $\mathbb{E}\epsilon^2$ равен некоторой фиксированной величине σ^2 , $\forall t_1 \neq t_2 \text{Cov}(\epsilon_{t_1}, \epsilon_{t_2}) = 0$. Очевидно, что белый шум является стационарным в широком смысле. Белый шум называется гауссовым, если для произвольной конечной последовательности моментов наблюдений t_1, \dots, t_n совокупность случайных величин $\epsilon_{t_1}, \dots, \epsilon_{t_n}$ подчиняется n -мерному нормальному распределению. Гауссов шум является стационарным также и в узком смысле.

Под процессом **случайного блуждания** понимается процесс, в котором $X_t = X_{t-1} + \epsilon_t$, где ϵ_t -белый шум. Предположим, что в момент времени $t = 0$ значение временного ряда, являющегося реализацией случайного процесса, составило x_0 . Очевидно, что значение процесса в момент времени T является случайной величиной, представимой в виде

$$X_T = x_0 + \sum_{t=1}^T \epsilon_t$$

Математическое ожидание случайной величины $X_T - x_0$ может быть представлено в виде

$$\mathbb{E}(X_T - x_0) = \mathbb{E}X_t - x_0 = \sum_{t=1}^T \epsilon_t = 0,$$

то есть $\mathbb{E}X_T = x_0$.

Дисперсия X_T очевидно равна

$$\mathbb{E}(X_t - x_0)^2 = \sum_{t=1}^T \sum_{t'=1}^T \epsilon_{t'} \epsilon_t$$

Предположим, что дисперсия белого шума равна σ^2 . Тогда очевидно

$$\mathbb{E}(X_T - x_0)^2 = \sum_{t=1}^T \epsilon_t^2 = T\sigma^2$$

Таким образом, дисперсия случайного блуждания в конце интервала пропорциональна длине этого интервала.

Процесс случайного блуждания очевидно является нестационарным.

Детерминированная и случайная компонента стохастического процесса

Любой стохастический процесс Y_t может быть представлен в виде $Y_t = D_t + X_t$, где D_t - детерминированная компонента, вычисляемая как функция от момента времени, X_t -случайная компонента.

Детерминированная компонента может быть представлена в виде суммы тренда R_t и сезонной компоненты S_t : $D_t = R_t + S_t$.

Линейный тренд: $R_t = \alpha + \beta t$, обычно выделяется на каком-то достаточно продолжительном интервале времени.

Сезонная компонента: $S(t) = A \cos(\omega t + \phi)$ представляет циклические изменения с фиксированным периодом. Процесс, в котором присутствует сезонная компонента стационарным не является

Теорема Вольда

Любой стационарный в ковариациях процесс может быть представлен в виде

$$X_t = \mu + \sum_{j=0}^{\infty} b_j \epsilon_{t-j},$$

где ϵ_t -белый шум, μ - математическое ожидание процесса. Следует отметить, что значения коэффициентов b_j зависят только от задержки j и не зависят от времени t . Сходимость по вероятности суммы

$$\sum_{j=0}^{\infty} b_j \epsilon_{t-j}$$

обеспечивается при существовании конечного предела $\sum_{j=1}^n |b_j|$ при $n \rightarrow \infty$

Процесс скользящего среднего

Стохастический процесс называется **процессом скользящего среднего** или moving average (MA(q)), если

$$X_t - \mu = \sum_{j=0}^q b_j \epsilon_{t-j},$$

где μ -математическое ожидание X_t . Название "скользящее среднее" связано с тем, что значение процесса в момент t является взвешенным средним по значениям белого шума в q предшествующих моментах времени. Математическое ожидание процесса MA(q) очевидно равно 0.

Дисперсия процесса MA(q) может быть выражена через сумму квадратов коэффициентов b_j

$$\mathbb{E}(X_t - \mu)^2 = \sigma^2 \sum_{j=0}^q b_j^2.$$

Очевидно, дисперсия процесса не зависит от t .

Ковариация случайных величин $x_t = X_t - \mu$ и $x_{t+\tau} = X_{t+\tau} - \mu$ может быть вычислена по формуле

$$\text{cov}(x_{t+\tau}, x_t) = \sigma^2 \sum_{j=0}^{q-\tau} b_j b_{j+\tau}$$

при $\tau \leq q$. При $\tau > q$ $\text{cov}(x_{t+\tau}, x_t) = 0$. Таким образом для процесса МА(q) удовлетворяются все три требования стационарности в ковариациях.

Случайный процесс называется **процессом авторегрессии**, если его значение в точке t является суммой белого шума в точке t и линейной комбинацией значений этого процесса в предшествующие моменты времени.

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \dots + a_p X_{t-p} + \epsilon_t$$

Процесс авторегрессии представленного вида обозначается AR(p). Вообще говоря процесс авторегрессии не обязательно является стационарным в ковариациях. При исследовании и прогнозировании временных рядов принято рассматривать объединённый процесс **авторегрессии-скользящего среднего**

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \dots + a_p X_{t-p} + b_1 \epsilon_{t-1} + b_2 \epsilon_{t-2} + \dots + b_q \epsilon_{t-q} + \epsilon_t$$

Эффективное прогнозирование временного ряда очевидно возможно при наличии в нём долговременных трендов или сезонных колебаний. Для того, чтобы использовать при прогнозировании наличие линейного тренда, достаточно найти по имеющемуся временному ряду коэффициенты линейной регрессионной модели $X_t = \alpha + \beta t + \epsilon$. Это можно сделать, например, с помощью метода наименьших квадратов. Учёт сезонных колебаний может быть осуществлён через подсчёт средних значений по множеству точек временного ряда, соответствующих одинаковой фазе цикла. В случае, если временной ряд является стационарным в ковариациях, для прогнозирования могут использоваться математические модели, основанные на процессе авторегрессии-скользящего среднего. Предварительным этапом, обеспечивающим успех моделирования, является переход от исходного необязательно стационарного ряда к стационарному.

Проверка стационарности может производиться через анализ автокорреляций с различной величиной лага. Автокорреляция с лагом k вычисляется как

$$r_k = \frac{\sum_{t=k+1}^T (X_t - \bar{X})(X_{t-k} - \bar{X})}{T\hat{D}},$$

где \bar{X} - среднее значение, \hat{D} - дисперсия X на интервале наблюдения. Если временной ряд является стационарным, то r_k должны убывать и приближаться к 0 с увеличением k . Если с ростом k сохраняются высокие значения r_k , то процесс нестационарен.

Тест Дикки-Фуллера. Случайный процесс может оказаться нестационарным вследствие того, что он является процессом случайного блуждания: $X_t = X_{t-1} + \epsilon_t$. Процесс случайного блуждания также называют процессом с единичным корнем в связи с тем, что одним из корней соответствующего характеристического уравнения является 1.

Процесс случайного случайного блуждания очевидно является частным случаем простого процесса авторегрессии

$$X_t = \rho X_{t-1} + \epsilon_t. \quad (8)$$

Уравнение (8) может быть записано в виде

$$\Delta_t = X_t - X_{t-1} = \gamma X_{t-1} + \epsilon_t$$

где $\gamma = \rho - 1$.

Процессу случайного блуждания соответствует $\gamma = 0$ или $\rho = 1$.

Проверка гипотезы о том, соответствует ли некоторый временной ряд процессу случайного блуждания или процессу с единичным корнем, может быть сведена к проверке гипотезы о равенстве 0 коэффициента γ .

Проверка равенства $\gamma = 0$ частично аналогична рассматриваемой ранее проверке нулевой гипотезы о равенстве 0 регрессионного коэффициента с помощью теста Стьюдента. Для проверки равенства $\gamma = 0$ используется та же самая статистика. Выполнение неравенств $\rho > 1$ и соответственно $\gamma > 0$ приводит к резкому "взрывному" возрастанию величин X за короткие промежутки времени. Подобные процессы крайне редко встречаются на практике. Поэтому на практике актуальной оказывается проверка гипотезы $\gamma = 0$ против альтернативы $\gamma < 0$. То есть используется односторонний критерий.

Переход к стационарным временным рядам может быть произведён с помощью взятия разностей двух следующих друг за другом элементов.

- Очевидно, что разности двух соседних элементов процесса случайного блуждания $\Delta X_t = X_t - X_{t-1} = \epsilon_t$ составляют стационарный ряд, являющийся белым шумом.
- Рассмотрим также нестационарный случайный процесс $X_t = \alpha + \beta t + \gamma t^2 + \epsilon_t$. Разность двух последующих элементов, которую мы назовём разностью первого порядка, может быть записана как

$$\Delta X_t = \alpha + \beta t + \gamma t^2 - \alpha - \beta(t-1) - \gamma(t-1)^2 + \epsilon_t - \epsilon_{t-1}$$

- Разность двух последующих элементов ряда, составленного из разностей первого порядка, (разность второго порядка) может быть записана в виде

$$\Delta^2 X_t = \Delta X_t - \Delta X_{t-1} = 2\gamma + \epsilon_t - 2\epsilon_{t-1} + \epsilon_t$$

Таким образом, ряд, составленный из разностей второго порядка является стационарным.

Последовательное взятие разностей увеличивающегося порядка является универсальным средством достижения стационарности.

Временной ряд называется интегрированным порядка k , если временной ряд, составленный из разностей порядка k ($\Delta^k X_t$), является стационарным. Наличие стационарности порядка k для ряда X_t обозначается $X_t \sim I(k)$. Временной ряд, удовлетворяющий условию $X_t \sim I(0)$ очевидно является стационарным. Процесс случайного блуждания является процессом $\sim I(1)$.

Прогнозирование временного ряда с использованием модели авторегрессии-скользящего среднего.

Будем проводить последовательное взятие разностей между соседними элементами временного ряда до тех пор, пока не определим порядок его интегрированности k .

Далее для прогнозирования временного ряда $\{Y_t = \Delta^k X_t | t = 1, \dots, m\}$ будем использовать модель авторегрессии-скользящего среднего:

$$Y_t = \sum_{i=1}^p a_i Y_{t-i} + \sum_{j=1}^q b_j \epsilon_{t-j} + \epsilon_t \quad (9)$$

Для того, чтобы можно было прогнозировать значение процесса в точке t очевидно необходимо знать величины $p, q, a_1, \dots, a_p, b_1, \dots, b_q$.

Прогнозирование временного ряда с использованием модели авторегрессии-скользящего среднего.

Поиск значений p, q проводится через изучение графика автокорреляционной функции (ACF) и частной автокорреляционной функции (PACF). ACF является функцией натурального аргумента и определяется точке i как коэффициент корреляции Пирсона между Y_t и Y_{t-i} : $\gamma(i) = \rho(Y_t, Y_{t-i})$. PACF также является функцией натурального аргумента и описывает линейную связь между Y_t и Y_{t-i} за вычетом влияния Y_t всех элементов ряда, находящихся между Y_t и Y_{t-i} . Значение PACF в точке i определяется как коэффициент корреляции Пирсона между величиной $Y_t - \sum_{j=1}^i \phi_j Y_{t-j}$, где

$$(\phi_1, \dots, \phi_{i-1}) = \arg \min \left[\sum_{t=i}^m \left(Y_t - \sum_{j=1}^{i-1} \phi_j Y_{t-j} \right)^2 \right].$$

Прогнозирование временного ряда с использованием модели авторегрессии-скользящего среднего.

Величины p и q подбираются из условия выхода на приблизительно нулевой уровень соответственно ACF $\gamma(i)$ и PACF $\gamma_p(i)$.

Предположим, что величины p, q заданы. Тогда для поиска оптимальных значений $a_1, \dots, a_p, b_1, \dots, b_q$ может быть использована следующая схема.

- Зададим некоторое некоторое конечное множество \tilde{B} векторов значений параметров b_1, \dots, b_q , внутри которого предполагается проводить поиск

Процесс авторегрессии-скользящего среднего. Поиск регрессионных параметров

- Зафиксируем некоторый $\mathbf{b}^* \in \tilde{B}$
- Ведём новые переменные

$$Z_1 = X_1, Z_2 = X_2 - b_1^* Z_1, \dots, Z_k = X_k - \sum_{i=1}^{q-1} b_i^* Z_{k-i}, \dots \quad (10)$$

- $\forall t$ справедливо

$$Z_t = \sum_{i=1}^p a_i Z_{t-i} + \epsilon_t. \quad (11)$$

Условия справедливости данного факта будет обсуждаться ниже. Поскольку все значения Z вычисляются по временному ряду, то коэффициенты a_1, \dots, a_p могут быть найдены исходя из требования минимизации

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{t=1}^n \left(Z_t - \sum_{i=1}^p a_i Z_{t-i} \right)^2$$

- Повторим два предыдущих пункта при различных значениях $\mathbf{b} \in \tilde{B}$ и выберем \mathbf{b} при котором $\sum_{t=1}^n \epsilon_t^2$

Справедливость равенства $Z_t = \sum_{i=1}^p a_i Z_{t-i} + \epsilon_t$ при произвольном t может быть показана через представление процессов AR(p) и MA(q) с использованием оператора сдвига L . Действие оператора L заключается в переходе от значения переменной в точке t к значению этой же переменной в точке $t - 1$: $LX_t = X_{t-1}$, $L\epsilon_t = \epsilon_{t-1}$.

Действие оператора L^k заключается в последовательном применении k раз оператора L : $L^k X_t = X_{t-k}$

Естественным образом вводится сложение операторов и умножения оператора на число: $aL^k X_t = aX_{t-k}$,
 $(a_1 L^{k_1} + a_2 L^{k_2})X_t = a_1 X_{t-k_1} + a_2 X_{t-k_2}$

Процесс AM(q) может быть представлен в виде операторного полинома

$$X_t = (1 + b_1 L + \dots + b_q L^q) \epsilon_t$$

Процесс авторегрессии очевидно может быть представлен в виде

$$X_t = (a_1L + a_2L^2 + \dots, a_pL^p)X_t + \epsilon_t$$

Процесс ARMA(p,q) может быть задан уравнением

$$(1 - a_1L - \dots - a_pL^p)X_t = (1 + b_1L + \dots + b_qL^q)\epsilon_t$$

Откуда следует, что

$$X_t = \frac{(1 + b_1L + \dots + b_qL^q)\epsilon_t}{1 - a_1L - \dots - a_pL^p}$$

Последнее справедливо только, если обратный оператор к оператору $1 - a_1L - \dots - a_pL^p$ действительно существует.

Пусть $Z_t = \frac{\epsilon_t}{1 - a_1 L - \dots - a_p L^p}$.

Тогда $(1 - a_1 L - \dots - a_p L^p)Z_t = \epsilon_t$, что эквивалентно (11)

$X_t = (1 + b_1 L + \dots + b_q L^q)Z_t$, что эквивалентно (10).

После того, как установлены величины p и q , а также коэффициенты a_1, \dots, a_p и b_1, \dots, b_q , необходимо оценить адекватность полученной модели.

Диагностика модели естественно должна производиться по величинам остатков. Из определения процесса ARMA следует, что величины $\epsilon_1, \dots, \epsilon_T$ являются белым шумом. Откуда следует независимость остатков и следовательно близость 0 значений автокорреляционной функции.

Для проверки правильности выбора сложности модели, то есть количеств слагаемых p и q принято использовать информационные критерии: информационный критерий Акаике (AIC) или байесовский информационный критерий Шварца (BIC).

Под многомерным временным рядом понимается ряд, содержащий в каждой временной точке значения нескольких переменных $\{X_t^1, \dots, X_t^k | t = 0, \dots, T\}$.

Наряду с задачами прогнозирования значений переменных, входящих в многомерный временной ряд, нередко возникают задачи изучения взаимосвязи этих переменных между собой. Например, линейная взаимосвязь может изучаться путём вычисления коэффициентов корреляции Пирсона для пар переменных. Влияние переменных X^{i_1}, \dots, X^{i_r} на переменную X^{i_y} исследуется через построение соответствующей регрессионной модели.

Вместе с тем многочисленные расчёты показывают относительно частую встречаемость ситуаций, когда

- коэффициент корреляции между двумя переменными X^i и X^j из одного и того же временного ряда имеет высокую статистическую значимость при оценивании последней при использовании стандартных средств верификации
- одномерные временные ряды, содержащие отдельно значения переменных X^i и X^j , остаются нестационарными даже после исключения из него тренда и сезонной периодичности
- наличие корреляции между X^i и X^j не поддаётся объяснению в рамках соответствующих научных дисциплин

Аналогичные ситуации нередко возникают и для более сложных регрессионных моделей, при тестировании последних с помощью стандартных статистических критериев.

Возрастание вероятности чисто случайного возникновения связей, формально оцениваемые как статистически достоверные с использованием стандартных статистических тестов, подтверждаются в экспериментах с искусственно сгенерированными данными.

Например, при многократном искусственном генерировании временных рядов значений переменных X^i и X^j с использованием процесса случайного блуждания при нормально распределённой случайной компоненте ϵ_t , доля рядов со значением коэффициента корреляции $\rho(X^i, X^j) \in C$, где C является критической областью двухстороннего критерия Стьюдента для уровня значимости α существенно превышает саму величину α .

С целью повышения достоверности регрессионного анализа на временным рядах при нестационарности целевой переменной Y_t и регрессоров X_t^1, \dots, X_t^k , Грэйнджером был предложено дополнительное требование стационарности остатков регрессионной модели $Y_t = \sum_{i=1}^k \beta_i X_t^i + r_t$. То есть вводилось требование стационарности ряда

$$Y_t - \sum_{i=1}^k \beta_i X_t^i. \quad (12)$$

Требование стационарности ряда (12) является тождественным требованию коинтеграции. Пусть $\{Z_t^1, \dots, Z_t^k\}$ является совокупностью интегрированных по первому порядку временных рядов, т.е. $Z_t^i \sim I(1), i = 1, \dots, k$

Временной ряд $\{Z_t^1, \dots, Z_t^k\}$ называется коинтегрированным, если существует такой вектор (a_1, \dots, a_k) , что временной ряд $u_t = \sum_{i=1}^k a_i Z_t^i$ является стационарным процессом, т.е. $u_t \sim I(0)$. Очевидно, что требование коинтеграции в последнем смысле является фактически эквивалентным условию стационарности остатков в смысле (12) из-за того, что свойство стационарности не теряется при умножении ряда на произвольную константу. Наличие коинтеграции между переменными говорит о существовании определённой связи между ними, несмотря на случайный непредсказуемый характер изменений каждой из них.