Лектор – *Сенько Олег Валентинович*

Курс «ПРИКЛАДНАЯ СТАТИСТИКА И АНАЛИЗ ДАННЫХ ЧАСТЬ II»
ПС

Перестановочный тест

Эффективным непараметрическим инструментом статистической верификации является перестановочный тест. Целью перестановочного теста является проверка существования зависимости между целевой переменной Y от и набором пременных $X_1, \ldots, X_n$ в рамках некоторой гипотезы о соответствующей математической модели. Проверка существования зависимости сводится к попытке опровержения нулевой гипотезы о независимости переменной Y от переменных $X_1, \ldots, X_n$ с использованием статистики, характеризующей качество аппроксимации зависимости с использование 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 $X_1, \ldots, X_n$ статистикой критерия может служить величина R^2.

Перестановочный тест

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

- Сделать предположение о характере зависимости и предложить соответствующую этому предположению статитику T.
- Рассчитать значение статитики T для исходной выборки $\widetilde{S}$
- Выбрать количество перестановок N исходя из имеющихся ревычислительных ресурсов.
- Приравнять 0 целочисленны показатель N_t, подсчитывающей достижение или привышение статистики критерия при справедливости нулевой гипотезы H_0 (см.далее)
- А) С использованием генератора случаных чисел получить случайную перестановк f чисел из набора $\{1, \ldots, m\}$

Перестановочный тест

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

Перестановочный тест.

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

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

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

Перестановочный тест. Обоснование.

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 $X_1, \ldots, X_n$ считаются детерминированными, то есть нвектора X-описаний для всех генерируемых выборок одинаковы и равны $\mathbf{x}_1, \ldots, \mathbf{x}_m$
- значениz Y генерируются случайно и независимо из маргинальных эмпирического распределения $\hat{E}_y$. Такая генерация совпадает с выборкой с возвращением.

Перестановочный тест. Обоснование.

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 $\widetilde{S}$ и распределению на ней значений Y

Theorem

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

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

Перестановочный тест. Обоснование.

Доказательство Обозначим через $W^p(\widetilde{S})$ множество выборок, генерируемых описанным выше процессом. Разобьём $W^p(\widetilde{S})$ на группы неразличимых выборок $\widetilde{G}_1, \ldots, \widetilde{G}_{N_d}$. Выборки $\widetilde{S}'$ и $\widetilde{S}''$ считаются тождественнными, 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 $\mathbf{x}-$частях, с помощью которых могут быть получены все выборки из группы является одинаковым для всех групп.
- 3)Вероятности появления каждой из групп одинаковы.

Перестановочный тест. Обоснование.

Доказательство пункта 1) очевидно.
Докажем пункт 2). Предположим,что Y в $\widetilde{S}$ принимает значения $k \leq m$ значений. Пусть $l_1, \ldots, l_k$ являются количествами объектов в $\widetilde{S}$, на которых Y принимает значения $\breve{y}_1, \ldots, \breve{y}_k$ соответственно. Любая перестановка, переводящая $\widetilde{S}$ 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 $\widetilde{\pi}_u$, где $\widetilde{\pi}_u$ является множеством перестановок, сохраняющих принадлежность G_u. Число перестановок в $\widetilde{\pi}_u$ не зависит от номера группы u и равна $T_g = \prod_{i=1}^{k} l_i!$.

Перестановочный тест. Обоснование.

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

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

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

Перестановочный тест. Преимущества и недостатки.

На правктике вместо подсчёта всевозможных перестановок используется их подмножество, получаемо 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 являетсянеобходимость больших объемов вычислений.

Проблема множественного тестирования

Целью многих исследований в различных областях науки является выяснение, какие показатели из некоторого заранее заданного набора $X_1, \ldots, X_n$ связаны с целевой переменной Y.
Стандартной процедурой устанавления 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 равнв α.

Проблема множественного тестирования

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 $X_1, \ldots, X_n$. 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 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 гипотезы.

Проблема множественного тестирования

Наряду с FWER для оценки множественноно 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 $\mathbb{E}(\frac{V}{R}|R > 0)P(R > 0)$

Проблема множественного тестирования

Аналогично тому, как для оценивания значимости отдельных связей проверяется неравенство $p \leq \alpha$, для оценки значимости с учётом эффекта множественного тестирования может проверяться неравенство $FWER \leq \alpha$. Попытаемся оценить вероятность случайного отвержения по крайней мере одной из нулевые гипотез $H_0^1, \ldots, H_0^n$ 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 выполняось 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 $\widetilde{p}_j = \min(np_j, 1)$.
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 $\frac{0.01}{n}$.

Проблема множественного тестирования. Метод Шидака

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

$$\widetilde{p}_j = 1 - (1 - p_j)^m$$

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

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

Неравенство Шидака справедливо в частности длч многомерных нормальных распределений.

Проблема множественного тестирования. Одношаговые процедуры

Недостатком методов Бонферрони и Шидака является учёт при коррекции на множественное только лишь статистической значимости, при которой отвергаются индивидуальные нулевые гипотезы. При этом не учитывается общее число отвергнутых на таком уровне гипотез. Подобные процедуры принято называть одношаговыми (single step). Одной из возможных пособов является подход Вестфолла и Янга (Westfall and Youg)в котором рассчитанные $p-$значения рассматриваются как реализации случайных величин $P_1, \ldots, P_n 1$ скорректированное значение определяется по формуле

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

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 вычиляться по формуле:

$$\widetilde{p}_i = P(\max_{1 \le l \le n} T_l \ge t_i | H_u). \tag{5}$$

Предполагается, что рассчитанные значения статистик критерия для каждой из нулевых гипотез $t_1, \ldots, t_n$ являются реализация случайных функций $T_1, \ldots . T_n$ соответственно.
Для вычисления скорректированных значений согласно формулам (4,5) может быть использован также перестановочный тест.

Проблема множественного тестирования. Перестановочный тест.

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

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

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

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

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

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

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

Проблема множественного тестирования. Метод Холма

Проводить учёт общего числа отвергнутых индивидуальных нулевых гипотез позволяют процедуры пошагового спуска (step down). Одной из таких процедур является метод метод Бонферрони-Холма.
Предположим, что для нулевых гипотез $H_0^1, \ldots, H_0^n$ рассчитаны $p-$значения $p_1, \ldots, p_m$. Пусть $p^{(1)}, \ldots, p^{(m)}$ - ряд упорядоченных по возрастанию $p-$значений и $H_0^{(1)}, \ldots, H_0^{(m)}$ - ряд соответствующих нулевых гипотез.
Зафиксируем уровень значимости α.

Проблема множественного тестирования. Метод Холма

Предположим, что h - минимальный индекс,удовлетворяющий условию:

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

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

Временные ряды.Основные понятия

По временным рядом понимается 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 соответсттвие 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 $\{X_t, t = \ldots, -2, -1, 0, 1, 2, \ldots\}$.
Очевидно, что стохастический процесс полностью характеризуется совокупностью плотностей совместных распределений случайных величин вида $f(x_{t_1}, \ldots, x_{t_n})$, где $t_1, \ldots, t_n-$ некоторый набор мометов наблюдений.

Строгая стационарность

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

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

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

Строгая стационарность

- 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
- Автоковариционная 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

$$Cov(X_{t_1}, X_{t_2}) = \int\int (x_{t_1} - \mu)(x_{t_2} - \mu) f(x_{t_1}, x_{t_2}) dx_{t_1} dx_{t_2} =$$

$$= \int\int (x_{t_1} - \mu)(x_{t_2} - \mu) f(x_{t_1+\Delta}, x_{t_2+\Delta}) dx_{t_1} dx_{t_2}$$

Слабая стационарность

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.

Примеры случайных процессов

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, $\forall t_1 \neq t_2\ Cov(\epsilon_{t_1}, \epsilon_{t_2}) = 0$. Очевидно, что белый шум является стационарным в широком смысле. Белый шум называется гауссовым, если для произвольной конечной последовательности моментов наблюдений $t_1, \ldots, t_n$ совокупность случайных величин $\epsilon_{t_1}, \ldots, \epsilon_{t_n}$ подчиняется $n-$мерному нормальному распределению. Гауссов шум является стационарным также и в узком смыле.

Примеры случайных процессов. Процесс случайного блуждания.

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

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

Процесс случайного блуждания.

Дисперсия $X_T - x_0$ 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.

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 фиксированыым периодом. Процесс, в котором присутствует сезонная компонента стационарным не являетсяю

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 $n \to \infty$

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

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

Процесс скользящего среднего

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

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

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

Процесс авторегрессии-скользящего среднего

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

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

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

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \ldots + a_p X_{t-p} + b_1 \epsilon_{t-1} + b_2 \epsilon_{t-2} + \ldots$$

$$+ b_q \epsilon_{t-q} + \epsilon_t$$

Процесс авторегрессии-скользящего среднего

Математические модели, основанные на гипотезе генерации временного ряда из процесса авторегрессии-скользящего среднего могут быть использованы для прогнозирования временного ряда. Для того, чтобы можно было прогнозировать значение процесса в точке t очевидно необходимо знать величины $p, q, a_1, \ldots, a_p, b_1, \ldots, b_q$. Предположим, что величины p, q заданы. Тогда для поиска оптимальных значений $a_1, \ldots, a_p, b_1, \ldots, b_q$ может быть использована следующая схема.

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

Процесс авторегрессии-скользящего среднего. Поиск регрессионных параметров

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

$$Z_1 = X_1, Z_2 = X_2 - b_1^* Z_1, \ldots, Z_k = X_k - \sum_{i=1}^{q-1} b_i^* Z_{k-i}, \ldots \tag{7}$$

- $\forall t$ справедливо

$$Z_t = \sum_{i=1}^{p} a_i Z_{t-i} + \epsilon_t. \tag{8}$$

Условия справедливости данного факта будет обсуждаться ниже. Поскольку все значения Z вычисляются по временному ряду, то коэффициенты $a_1, \ldots, a_p$ могут быть найдеы исходя из требования минимизции

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

Операторная форма представления случайного процесса

- Повторим два предыдущих пункта при различных значениях $\mathbf{b} \in \widetilde{B}$ 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

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

Операторная форма представления случайного процесса

Процесс авторегрессии очевидно может быть представлен в виде

$$X_t = (a_1 L + a_2 L^2 + \ldots, a_p L^p) X_t + \epsilon_t$$

Процесс ARMA(p,q) может быть задан уравнением

$$(1 - a_1 L - \ldots - a_p L^p) X_t = (1 + b_1 L + \ldots + b_q L^q) \epsilon_t$$

Откуда следует, что

$$X_t = \frac{(1 + b_1 L + \ldots + b_q L^q) \epsilon_t}{1 - a_1 L - \ldots - a_p L^p}$$

Последнее справедливо только, если обратный оператор к оператору $1 - a_1 L - \cdots - a_p L^p$ действительно существует.

Поиск параметров ARMA(p,q)

Пусть $Z_t = \frac{\epsilon_t}{1 - a_1 L - \ldots - a_p L^p}$.
Тогда $(1 - a_1 L - \ldots - a_p L^p) Z_t = \epsilon_t$, что эквивалентно (8)
$X_t = (1 + b_1 L + \ldots + b_q L^q) Z_t$, что эквивалентно (7).