ПРИМЕНЕНИЕ РЕГРЕССИОННОГО АНАЛИЗА В ЗАДАЧАХ ТЕОРИИ ТЕЛЕТРАФИКА

DOI: 10.36724/2072-8735-2020-14-12-18-25

Manuscript received 20 July 2020;
Accepted 28 September 2020

Ключевые слова: *регрессионный анализ, метод наименьших квадратов, синусоидальная модель, полиномиальная модель, прогнозирование, изменение данных, предиктор, оценка, машинное обучение, статистические данные, измерения,интеллектуальный анализ данных*

Шерстнева Алина Анатольевна,
Сибирский Государственный Университет Телекоммуникаций и Информатики,
г. Новосибирск, Россия, ***asherstneva@sibguti.ru***

Рассматривается метод наименьших квадратов для решения задач теории систем массового обслуживания. Показана возможность прогнозирования поведения инфокоммуникационной системы и выбора оптимальной модели ее функционирования. В качестве информационной основы его применения взяты статистические данные мониторинга инфокоммуникационных систем. Целью является расчет параметров оптимальной модели тренда, характеризующей тенденцию развития случайных процессов во времени. Для получения результатов максимально приближенных к реальным значениям показателей функционирования инфокоммуникационных систем рассмотрены полиномиальная и синусоидальная модели. Предлагается использовать метод регрессионного анализа для определения значений параметров для функции по набору данных наблюдений. В теоретических исследованиях также приведено использование линейного и нелинейного метода наименьших квадратов применительно к окружности. Задача экспериментального анализа заключается в получении оценки параметров синусоидальной, полиномиальной моделей и центра окружности. Экспериментальный анализ выполнен с помощью программы математического моделирования Matlab. Сгенерирована равномерно распределенная случайная последовательность и случайная последовательность с нормальным распределением. Рассчитана последовательность с экспериментальными данными соответственно синусоидальной и полиномиальной моделей. В графическом виде показано соответствие модели для сгенерированных данных. Данные измерения подчиняются синусоидальной модели, последовательность измерений подчиняется полиномиальной модели. Расчетные параметры сведены в таблицу. Выполнена оценка порядка многочлена. Получена расчетная дисперсионная кривая полиномиальной модели. Приведены расчетные значения дисперсии полиномиальной модели. Сделана оценка данных измерений. Оценка показывает значения достаточно близкие к реальным данным. Результаты приведены на графиках. Расчетные коэффициенты достаточно близки по значениям к реальным коэффициентам полиномиальной модели. Также в графическом виде приведена примерная модель окружности данных измерений. Получены достаточно близкие значения центра окружности и радиуса.

Информация об авторе:
Шерстнева Алина Анатольевна, *к.т.н., Сибирский Государственный Университет Телекоммуникаций и Информатики, г. Новосибирск, Россия*

Для цитирования:
Шерстнева А.А. Применение регрессионного анализа в задачах теории телетрафика // T-Comm: Телекоммуникации и транспорт. 2020. Том 14. №12. С. 18-25.

For citation:
Sherstneva A.A. (2020) Regression analysis application for teletraffic theory tasks. *T-Comm*, vol. 14, no.12, pp. 18-25. (*in Russian*)

Введение

При проектировании и дальнейшей эксплуатации инфокоммуникационных систем решается ряд задач, связанных с обеспечением их «жизненного цикла». Круг решаемых задач достаточно широк. Однако, прежде всего, это задачи обеспечения надежности и работоспособности.

Задачи формулируются как задачи теории телетрафика, теории систем массового обслуживания. Инфокоммуникационные системы/сети представляются в виде математических моделей, которые в свою очередь представлены графом состояний. Каждое состояние символизирует нахождение системы на определенном этапе обслуживания вызовов/заявок. Формируется ряд входных параметров, символизирующих переход из одного состояния системы в другое. В большинстве случаев входные параметры представляются независимыми случайными величинами с определенным видом распределения, как правило, экспоненциальным. Что, конечно, не всегда соответствует действительности. На выходе стремятся получить показатели, характеризующие определенные свойства рассматриваемой системы. Например надежность, работоспособность, безотказность, масштабируемость, ремонтопригодность и многие другие.

Существуют расчетные, экспериментальные, расчетно-экспериментальные методы определения выходных показателей. Ряд показателей можно определять по данным наблюдений за работой системы в процессе ее эксплуатации. Ряд показателей возможно определять только с помощью расчетных (теоретических) методов. Но и в том, и в другом случае для получения показателей максимально приближенных к реальным данным, в состав расчетных формул должны входить статистические данные, полученные системой мониторинга.

Современные системы мониторинга обладают способностью собирать и обрабатывать большие объемы статистических данных практически за любой период времени. Выборка данных может быть полностью определенной или не полностью определенной. С большой долей вероятности полученные вероятностно-временные характеристики работоспособности инфокоммуникационной системы будут соответствовать реальному, текущему положению. Но интерес представляет и прогнозирование этих показателей, например, при изменении условий эксплуатации, масштабировании системы/сети, сезонном увеличении пиковых нагрузок, изменении потоковой маршрутизации, периодически возникающих пограничных ситуациях, связанных с выходом из рабочей конфигурации отдельных системных/сетевых элементов. Решением задачи прогнозирования является составление математической модели функционирования инфокоммуникационной системы с учетом все вышеперечисленных факторов. При этом число исходных параметров будет значительно превышать число результирующих. Например, для оценки надежности системы рассчитывают коэффициент готовности. Но в расчетную формулу входит целый ряд исходных параметров, таких как интенсивность отказов, интенсивность поступления заявок на обслуживание, число обслуженных/потерянных заявок и многие другие. Исходные параметры являются случайными величинами, спрогнозировать которые достаточно трудно. Для этого необходимо большое число наблюдений с вводом определенных одного или нескольких критериев. Кроме того, расчет результирующих параметров в математических моделях также зависит и от вида распределения исходных параметров. Не всегда экспоненциальное распределение исходных параметров соответствует реальной картине происходящих в инфокоммуникационных системах процессов. Таким образом, составляется статистическая модель. Любая статистическая модель должна быть подвергнута соответствующей проверке. Результатом проверки является получение количественных переменных, характеризующих, например, процесс обслуживания вызовов или производительность системы в целом или продуктивность работы сотрудников компании.

В статье предлагается использовать метод регрессионного анализа для определения значений параметров для функции по набору данных наблюдений. Рассматриваются полиномиальная и синусоидальная модели. Математическая модель процесса представляется полиномом, коэффициенты которого определяются методом наименьших квадратов. При рассмотрении синусоидальной модели можно опираться на нелинейный метод наименьших квадратов.

Для получения результатов, максимально приближенных к реальным значениям показателей инфокоммуникационных систем/сетей использование полиномиальной модели позволит повысить порядок полинома, тем самым улучшить аппроксимацию. А также приводит к линейной системе нормальных уравнений при определении коэффициентов уравнения регрессии методом наименьших квадратов.

При наличии двух или более предикторных переменных модель называется моделью множественной регрессии:

$$y_i = a_0 + a_1 x_{1,i} + a_1 x_{2,i} + ... + a_p x_{p,i} + z_i ,$$

где y – прогнозируемая переменная, а $x_1,...,x_k - k$ переменных-предикторов.

Каждая из переменных предиктора должна иметь численное значение. Коэффициенты $a_1,...,a_p$ измеряют влияние каждого предиктора после учета влияния всех других предикторов в модели. Таким образом, коэффициенты измеряют предельные эффекты переменных предиктора. Построение модели множественной линейной регрессии может потенциально генерировать более точные прогнозы, поскольку прогнозируемая переменная будет зависеть от нескольких предикторов и от влияния каждого из них. В статье рассматривается синусоидальная и полиномиальная регрессия [1, 2].

На практике, конечно, у нас есть набор наблюдений, но мы не знаем значений коэффициентов $a_1,...,a_p$. Они должны быть оценены на основе данных. Принцип наименьших квадратов обеспечивает способ эффективного выбора коэффициентов путем минимизации суммы квадратов ошибок. Поиск наиболее подходящих оценок коэффициентов называется «подгонкой» модели к данным или «обучением» модели. Метод регрессионного анализа является видом машинного обучения, использование которого востребовано и актуально в настоящее время.

В статье при ссылке на оценочные значения используется обозначение $\hat{a}_1,...\hat{a}_1$.

Теоретические исследования

Полиномиальная модель

Критерием оценки метода наименьших квадратов является минимизация суммы квадратов отклонений (ошибок, для регрессионных моделей их часто называют остатками регрессии) между экспериментальными данными y_i и функцией $f(x_i)$ [1-5]:

$$\sum_{i=1}^{n} r_i^2 \text{ при } r_i = y_i - f(x_i)$$

Каждое наблюдение y_i состоит из систематической или объясняемой части модели, $a_0 + a_1 x_i$, и некоторой случайной «ошибки», z_i. Термин «ошибка» означает не ошибку, а отклонение от базовой модели прямой линии. В него входит все то, что может повлиять на y_i, кроме x_i. Для создания интервалов прогнозирования, принято, что ошибки имеют нормальное распределение с постоянной дисперсией σ^2. Когда используется модель линейной регрессии допускается, что каждый предиктор x не является случайно распределенной величиной. По данным наблюдений невозможно управлять значениями x, поэтому сделано предположение о виде распределения переменных [1, 2, 6-8].

Предполагая, что экспериментальные данные представляют собой полиномиальную функцию

$$y_i = a_0 + a_1 x_i + a_1 x_i^2 + ... + a_p x_i^p + z_i,$$

получаем выражение:

$$q(a_0, a_1, a_2, ..., a_p) = \sum_{i=1}^{n} r_i^2 = (y_i - (a_0 + a_1 x_i + a_1 x_i^2 + ... + a_p x_i^p))^2 \cdot$$

Далее составляется матрица для определения суммы квадратов ошибок и для $n \geq p+1$, получаем:

$$q(\Theta) = \sum_{i=1}^{n} r_1^2 = (y - H\Theta)^T (y - H\Theta).$$

Минимальное количество наблюдений $n = p+1$ для решения системы уравнений, которая привела бы к интерполяции экспериментальных данных. Поэтому для оценки методом наименьших квадратов требуется $n > p+1$ наблюдений. Параметр a_i, который минимизирует сумму квадратов ошибок, необходимо вычислить $p+1$ раз [1, 2].

Используя векторную запись q (Θ):

$$q(\Theta) = y^T y - 2y^T H\Theta + \Theta^T H^T H\Theta.$$

Для минимизации должно быть выполнено следующее необходимое условие $\nabla_g(\Theta) \overset{!}{=} 0$

Тогда:

$$\frac{\partial q(\Theta)}{\partial \Theta_0} = -2y^T H \underbrace{\frac{\partial \Theta}{\partial \Theta_0}}_{e_1} + \underbrace{\frac{\partial \Theta}{\partial \Theta_0}}_{e_1^T} H^T H\Theta + \Theta^T H^T H \underbrace{\frac{\partial \Theta}{\partial \Theta_0}}_{e_1},$$

$$\frac{\partial q(\Theta)}{\partial \Theta_p} = -2y^T H \underbrace{\frac{\partial \Theta}{\partial \Theta_p}}_{e_p} + \underbrace{\frac{\partial \Theta}{\partial \Theta_p}}_{e_p^T} H^T H\Theta + \Theta^T H^T H \underbrace{\frac{\partial \Theta}{\partial \Theta_p}}_{e_p},$$

Со столбцами e_i при $i = 1,2,..., p$ единичной матрицы I.

При дальнейшем упрощении получаем:

$$\frac{\partial q(\Theta)}{\partial \Theta_0} = \underbrace{-2y^T He_1}_{-2e_1^T H^T y} + \underbrace{e_1^T H^T H\Theta + \Theta^T H^T He_1}_{2e_1 H^T H\Theta},$$

$$\frac{\partial q(\Theta)}{\partial \Theta_p} = \underbrace{-2y^T He_p}_{-2e_p^T H^T y} + \underbrace{e_p^T H^T H\Theta + \Theta^T H^T He_p}_{2e_p H^T H\Theta} \cdot$$

Выражение для

$$\nabla_g(\Theta) = -2IH^T y + 2IH^T H\Theta = -2H^T y + 2H^T H\Theta \overset{!}{=} 0,$$

$$\Rightarrow H^T y = H^T \Theta$$

$$\hat{\Theta} = \underbrace{(H^T)^{-1} H^T}_{H+} y$$

Сумма квадратов ошибок:

$$q(\hat{\Theta}) = (y - H\hat{\Theta})^T (y - H\hat{\Theta}) = \left(y - H\underbrace{(H^T H)^{-1} H^T y}_{\hat{\Theta}}\right)^T \left(y - H\underbrace{(H^T H)^{-1} H^T y}_{\hat{\Theta}}\right) \cdot$$

Матрица проекции, также известная как матрица влияния, отображает вектор значений отклика (зависимых переменных) к вектору прогнозируемых значений:

$$P = H(H^T H)^{-1} H^T.$$

Матрица ортогональной проекции $P^{\perp} = (I - P)$ имеет ряд свойств:

- $P \cdot P = P$;
- $P = P^T$;
- $P^{\perp} \cdot P^{\perp} = P^{\perp}$;
- $P \cdot P^{\perp} = 0$;
- $(P^{\perp})^T = P^{\perp}$.

С учетом $Iy = y$:

$$q(\hat{\Theta}) = (Iy - Py)^T (Iy - Py);$$

$$q(\hat{\Theta}) = ((I - Py)^T)((I - P)y).$$

Применяя матрицу влияния и ее свойства:

$$q(\hat{\Theta}) = (P^{\perp} y)^T P^{\perp} y = y^T P^{\perp} P^{\perp} y = y^T (I - P) y.$$

$q(\hat{\Theta})$ является скаляром, применяя правило $tr(AB) = tr(BA)$:

$$q(\hat{\Theta}) = tr(y^T (I - P) y) = tr((I - P) yy^T).$$

Математическое ожидание:

$$E(q(\hat{\Theta})) = tr\left((I - P) E(yy^T)\right) = (n - (p+1))\sigma^2.$$

Выражение для определения оценки дисперсии:

$$\sigma^2 = \frac{E(q(\hat{\Theta}))}{n - (p+1)}; \quad \hat{\sigma}^2 = \frac{q(\hat{\Theta})}{n - (p+1)},$$

где n – количество наблюдений, и $p+1$ – количество параметров для оценки.

Синусоидальная модель

Экспериментальные данные для синусоидальной модели задаются выражением $y_i = a\sin(x_i + b) + z_i$ для определения параметров a и b таким образом, чтобы сумма квадратов отклонений между экспериментальными данными y_i и функцией $f(x_i)$ была минимальной [1, 2, 8]:

$$\sin(x+y) = \sin(x)\cos(y) + \cos(x)\sin(y);$$
$$f(x_i) = a\sin(x_i)\cos(b) + a\cos(x_i)\sin(b).$$

Поскольку, $A = a\cos(b)$ и $B = a\sin(b)$:

$$f(x_i) = A\sin(x_i) + B\cos(x_i);$$
$$q(a,b) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i(A\sin(x_i) + B\cos(x_i)))^2;$$
$$q(a,b) = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - ax_i^b)^2.$$

В матричной форме система линейных уравнений с неизвестными параметрами A и B принимает вид:

$$\begin{bmatrix} r_i \\ \vdots \\ r_n \end{bmatrix} = \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{y} - \underbrace{\begin{bmatrix} \sin(x_1) & \cos(x_1) \\ \vdots & \vdots \\ \sin(x_n) & \cos(x_n) \end{bmatrix}}_{H} \underbrace{\begin{bmatrix} A \\ B \end{bmatrix}}_{\Theta};$$

$$q(a,b) = \sum_{i=1}^{n} (y_i - A\sin(x_i) + B\cos(x_i))^2 = (y - H\Theta)^T (y - H\Theta).$$

Пользуясь методом наименьших квадратов для случая линейной регрессии из полиномиальной модели получаем:

$$\hat{\Theta} = (H^T H)^{-1} H^T y;$$
$$\hat{\Theta} = \begin{bmatrix} \hat{A} \\ \hat{B} \end{bmatrix}.$$

И для синусоидальной модели получаем:

$$\hat{a} = \sqrt{\hat{A}^2 + \hat{B}^2};$$
$$\hat{b} = \arctan\left(\frac{\hat{B}}{\hat{A}}\right).$$

Поскольку есть два неизвестных параметра, то необходимо два наблюдения для интерполяции – решения системы уравнений, и не менее трех для оценки МНК. Для полиномиальной модели требуется вычислить $p+1$ параметр, поэтому необходимо рассчитать два параметра для синусоидальной модели. Выражение для оценки дисперсии для синусоидальной модели принимает вид:

$$\hat{\sigma}^2 = \frac{q(\hat{\Theta})}{n-(p+1)} = \frac{q(\hat{\Theta})}{n-2}.$$

Нелинейный МНК применительно к окружности

Нелинейный МНК применим для задач, в которых при обработке экспериментальных, статистических данных используется формула, нелинейно зависимая от определяемых результирующих параметров. Такие задачи нередко встречаются в теории телетрафика при рассмотрении разных систем массового обслуживания.

Разница между центром окружности x_0 и точкой измерения x_k является радиусом окружности согласно экспериментальным данным [9, 10]:

$$x_k = \sqrt{(x_k - x_0)^2 + (y_k - y_0)^2} = r + z_k,$$

где r – истинный радиус окружности, а z_k – погрешность каждой точки измерения. Поэтому погрешность оценки можно записать как:

$$r_k - r = z_k.$$

Согласно методу МНК, сумма квадрата ошибки определяется:

$$q(x_0, y_0, r) = \sum_k \left(\sqrt{(x_k - x_0)^2 + (y_k - y_0)^2} - r\right)^2.$$

С требуемым минимумом:

$$\nabla_q = \begin{bmatrix} \frac{\partial q}{\partial x_0} \\ \frac{\partial q}{\partial y_0} \\ \frac{\partial q}{\partial r} \end{bmatrix} \overset{!}{=} 0.$$

Из приведенного выражения следует, что градиент приводит к нелинейной системе уравнений, которая может быть решена только численно, например, методом Ньютона.

Линейный МНК применительно к окружности

Нелинейная задача МНК подхода к окружности может быть линеаризована с помощью выражения Тейлора. Необходимо выполнить аппроксимацию центра окружности $(\hat{x}_0; \hat{y}_0)$ для вычисления радиуса r_k [9, 10]:

$$r_k(x_0, y_0) \approx r_k(\tilde{x}_0, \tilde{y}_0) + \frac{\partial r_k(x_0, y_0)}{\partial x_0}\bigg|_{\tilde{x}_0, \tilde{y}_0} (x_0 - \tilde{x}_0) + \frac{\partial r_k(x_0, y_0)}{\partial y_0}\bigg|_{\tilde{x}_0, \tilde{y}_0} (y_0 - \tilde{y}_0)$$
$$\approx r_k(\tilde{x}_0, \tilde{y}_0) + \frac{(x_k - \tilde{x}_0)}{r_k(\tilde{x}_0, \tilde{y}_0)}(x_0 - \tilde{x}_0) + \frac{(y_k - \tilde{y}_0)}{r_k(\tilde{x}_0, \tilde{y}_0)}(y_0 - \tilde{y}_0)$$

Новая модель для вычисления ошибки линейна в точке x_0, y_0:

$$z_k = r_k - r = r_k(\tilde{x}_0, \tilde{y}_0) + \frac{(x_k - \tilde{x}_0)}{r_k(\tilde{x}_0, \tilde{y}_0)}(x_0 - \tilde{x}_0) + \frac{(y_k - \tilde{y}_0)}{r_k(\tilde{x}_0, \tilde{y}_0)}(y_0 - \tilde{y}_0) - r.$$

Сумма квадратов ошибок определяется:

$$q(x_0, y_0, r) = \sum_k \left(c_k + a_k x + b_k y - r\right)^2 = (c - H\Theta)^T (c - H\Theta).$$

В матричной форме:

$$q(x_0, y_0, r) = \underbrace{\begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}}_{c} - \underbrace{\begin{bmatrix} a_1 & b_1 & 1 \\ \vdots & \vdots & \vdots \\ a_n & b_n & 1 \end{bmatrix}}_{H} \underbrace{\begin{bmatrix} x \\ y \\ r \end{bmatrix}}_{\Theta}.$$

Тогда оценка МНК для параметров:

$$\hat{\Theta} = (H^T H)^{-1} H^T c .$$

Также возможно улучшить оценку, повторяя расчет с определенными параметрами таким образом, чтобы сумма ошибок с новыми параметрами уменьшалась:

$$\hat{\Theta}(\tilde{x}_0, \tilde{y}_0) \to \hat{x}_0, \hat{y}_0$$
$$\hat{\Theta}(\hat{x}_0, \hat{y}_0) \to \bar{x}_0, \bar{y}_0$$
$$q(\tilde{x}_0, \tilde{y}_0, \tilde{r}) > q(\hat{x}_0, \hat{y}_0, \hat{r}) > q(\bar{x}_0, \bar{y}_0, \bar{r})$$

Условием для прекращения итераций станет:

$$\begin{bmatrix} \tilde{x}_0 \\ \tilde{y}_0 \\ \tilde{r} \end{bmatrix} - \begin{bmatrix} \bar{x}_0 \\ \bar{y}_0 \\ \bar{r} \end{bmatrix} < \delta \text{ или } q(\hat{x}_0, \hat{y}_0, \hat{r}) - q(\bar{x}_0, \bar{y}_0, \bar{r}) < e .$$

Экспериментальный анализ

Задача экспериментального анализа заключается в получении МНК-оценки синусоидальной, полиномиальной модели и центра окружности. Первоначально в программе Matlab была сгенерирована равномерно распределенная случайная последовательность x_1 на интервале [0,1] и случайная последовательность z_1 с нормальным распределением, каждая длиной 100 значений.

Равномерно распределенная последовательность $x = (x_1, x_2, x_3, ..., x_{100})^T$ на интервале [0,1] для синусоидальной модели (x_1), полиномиальной модели (x_2) и окружности (x_3) определяет следующие последовательности:

$$x_1 = x \cdot 4\pi ;$$
$$x_2 = x \cdot 5 ;$$
$$x_3 = z_3 \cos(x \cdot 2\pi) + 4.$$

Нормально распределенная случайная последовательность $z = (z_1, z_2, z_3, ..., z_{100})^T$ представляет собой следующие уравнения:

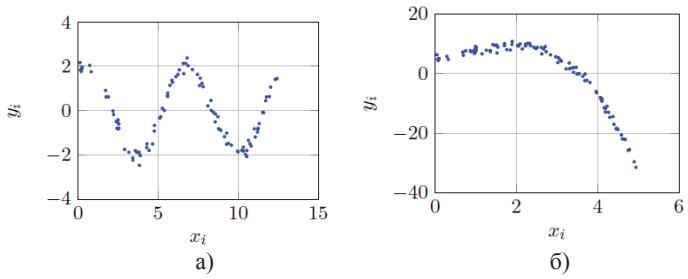
$$z_1 = z\sqrt{0{,}05} ;$$
$$z_2 = z ;$$
$$z_3 = z \cdot 0{,}05 + 6 + 4.$$

Далее рассчитывается последовательность $y = (y_1, y_2, y_3, ..., y_{100})^T$ с экспериментальными данными соответственно синусоидальной и полиномиальной моделей:

$$y_1 = 2\sin(x_1 + 1) + z_1 ;$$
$$y_2 = -0{,}6x_2^3 + 0{,}9x_2^2 + 3x_2^1 + z_2 + 4{,}5 + z_2 ;$$
$$y_3 = z_3 \sin(x \cdot 2\pi) + 2.$$

В Matlab последовательности x и y для каждого измерения сохраняются в матрице xy с размерностью 100×2, идентифицируемой соответствующим индексом, т. е. xy_1, xy_2 и xy_3.

Векторы двух столбцов каждой матрицы соответствуют наблюдениям x_i и y_i конкретной модели. Данные измерения xy_1 подчиняются синусоидальной модели (рис. 1а), последовательность измерений в xy_2 подчиняется полиномиальной модели (рис. 1б).



а) б)

Рис. 1. Соответствие модели для сгенерированных данных: а) синусоидальная модель; б) полиномиальная модель

МНК-оценка моделей заключается в оценке параметров a, b и дисперсии отклонений измерений σ_z^2 линеаризованной модели. Для этого была написана функция *LSE*, которая работает с моделью.

Рассчитанные последовательности данных измерений соответствуют реальным значениям *a и b*.

Синусоидальная модель $y_1 = 2\sin(x_1 + 1) + z_1 \Rightarrow a = 2, b = 1$.

В таблице 1 приведены расчетные параметры модели.

Таблица 1

Расчетные параметры модели

Параметр	Синусоидальная модели
a	
â	2.0155
b	
b̂	1.0084
$\hat{\sigma}_z^2$	0.0379

Программная реализация этой задачи осуществляется в Matlab, результат приведен в табл. 1. МНК-оценка значений параметров модели является хорошим приближением к реальным значениям параметров.

Оценка порядка многочлена

Теперь задача состоит в оценке порядка p полиномиальной модели. Для этого необходимо рассчитать оценочную дисперсию в зависимости от порядка модели $p = 1,2,...,10$. Порядок, обеспечивающий наименьшую дисперсию, будет верным.

Согласно оценке, дисперсия уменьшается до порядка 3. После этого дисперсия практически постоянна с небольшим локальным максимумом для порядка 6 и монотонно уменьшается от 6 до 10 порядка. Поэтому полиномиальный порядок 3 представляется хорошим приближением к экспериментальным данным.

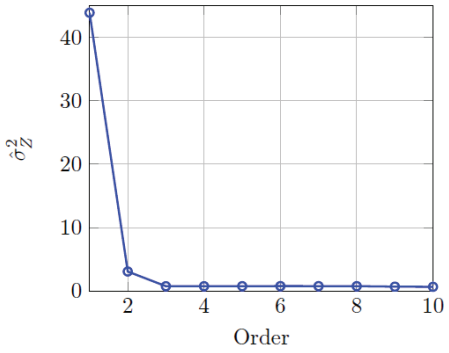


Рис. 2. Расчетная дисперсионная кривая полиномиальной модели с порядком p

Таблица 2

Расчетные значения дисперсии полиномиальной модели с порядком p

Порядок p	$\hat{\sigma}_z^2$
1	43.876
2	3.083
3	0.767
4	0.769
5	0.777
6	0.785
7	0.783
8	0.777
9	0.699
10	0.67

Таблица 3

Расчетные коэффициенты полиномиальной модели с порядком $p = 3$

Коэффициент	Расчетное значение	Реальное значение	Ошибка
a_0	4.7091	4.5	+0.2091
a_1	2.4780	3	-0.5220
a_2	1.1874	0.9	+0.2874
a_3	-0.6402	-0.6	-0.0402

Согласно МНК-оценке полиномиальной модели с порядком 3 расчетные коэффициенты достаточно близки по значениям к реальным коэффициентам полиномиальной модели.

МНК для определения центра окружности

Координаты местоположения центра x и y на рис. 3 можно оценить с помощью среднего арифметического:

$$\tilde{x}_0 = \frac{1}{N}\sum_k x_k \,;\; \tilde{y}_0 = \frac{1}{N}\sum_k y_k \,.$$

Данные измерений равномерно распределены по окружности, следовательно, среднее арифметическое является хорошей оценкой.

Для линейного МНК решение единственно. В нелинейном МНК решение необходимо находить с несколькими итерациями и решение будет зависеть от выбора начальной точки.

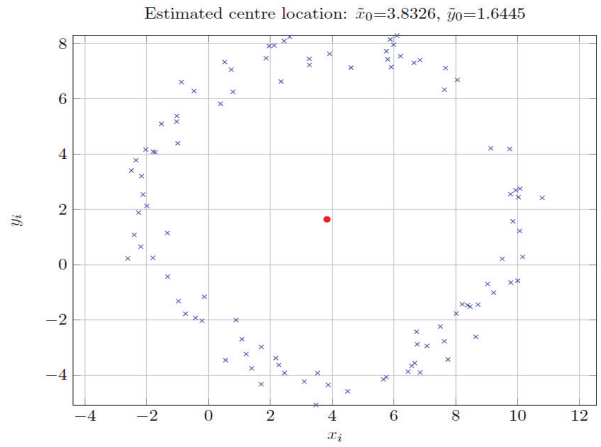


Рис. 3. Ориентировочное начальное расположение центра окружности

Далее выполняем оценку МНК, исходя из определенной центральной точки в качестве улучшенного начального значения.

МНК оценка повторяется до выполнения условия:

$$\left\|\tilde{x}_0(k), \tilde{y}_0(k) - \tilde{x}_0(k-1), \tilde{y}_0(k-1)\right\| < 10^{-10},$$

где $\tilde{x}_0(k), \tilde{y}_0(k)$ обозначает k-ую МНК-оценку центральной точки.

Восстановленная окружность и точки измерения приведены на графике рис. 4.

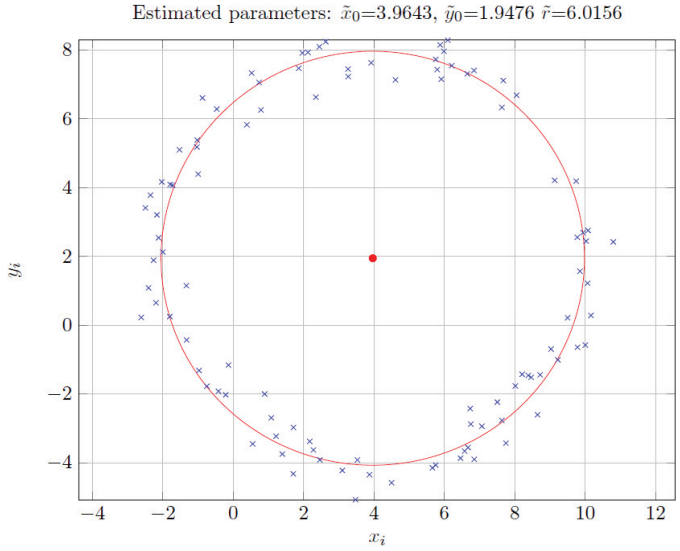


Рис. 4. Примерная модель окружности данных измерений

После нескольких итераций для оценки центра окружности по среднему арифметическому новый центр окружности определяется $\tilde{x}_0 = 3б9643$, $\tilde{y}_0 = 1{,}9476$, $\tilde{r}_0 = 6{,}0156$:

$$x_3 = z_3 \cos(x \cdot 2\pi) + 4\,;$$

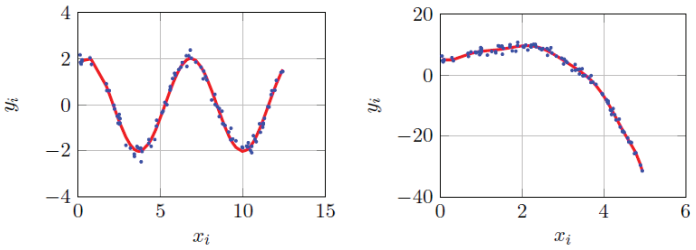
$$y_3 = z_3 \sin(x \cdot 2\pi) + 2\,.$$

Согласно сгенерированным данным измерений и графику на рис. 4, получаем достаточно близкие значения центра окружности и радиуса.

Заключение

В статье рассмотрены две модели: синусоидальная и полиномиальная. Для каждой из рассчитанных моделей приведены экспериментальные данные $(x_i, y_i) : i = 1,2,...,N$ и показана соответствующая реконструированная кривая $(x_i g(x_i)) : i = 1,2,...,N$.

На рисунке 3 экспериментальные данные выделены синим цветом, рассчитанные значения – красной линией. На рис. 3а показана кривая для синусоидальной модели, на рис. 3б показана кривая для полиномиальной модели.



а) б)

Рис. 5. МНК-оценка данных измерений:
а) синусоидальная модель; б) полиномиальная модель

Программная реализация этой задачи осуществлялась в Matlab. Из графиков рис. 5а и 5б видно, что МНК-оценка показывает значения достаточно близкие к реальным данным.

Применяемый метод математического программирования для полиномиальной и синусоидальной моделей делает возможным исключение недостатков классического регрессионного анализа.

Рассмотрен линейный и нелинейный метод наименьших квадратов применительно к окружности. Приведен МНК для определения центра окружности. В графическом изображении дана МНК-оценка данных измерений и примерная модель окружности данных измерений [9-14]. Регрессия широко используется в машинном обучении, типичным представителем алгоритма которого является линейная регрессия.

Применение машинного обучения полезно для прогнозирования и экстраполяции данных на новый формат математической модели. Экстраполяция данных подразумевает учет опыта предыдущих условий эксплуатации инфокоммуникационных систем/сетей в новых условиях, с новыми или дополненными функциональными возможностями.

Для обучения первоначально необходимо ранжировать исходные статистические данные по их значимости в зависимости от целевой переменной. Определиться среди множества статистических данных, какие будут зависимыми, а какие независимыми переменными. Как правило, прогнозируемой является зависимая переменная. Здесь и пригодиться практический опыт. Целевая переменная будет конечным результатом проводимых исследований. Далее выясняются аналитические зависимости между вероятностно-временными характеристиками и количественными показателями (целевой переменной). Таким образом, получается, что строится новая модель, в которую вводятся новые исходные статистические данные и вычисляется неизвестная прогнозируемая целевая переменная.

Для прогнозирования значимых характеристик инфокоммуникационных систем/сетей часто используют нейронные сети, как вид машинного обучения. Но в этом случае необходим большой объем исходных данных. И если с числом статистических данных, как правило, не возникает проблем, то разработка математической модели вызывает определенные затруднения. Для получения конкретного пригодного для практического использования результата, при составлении модели вводится ряд допущений, которые в итоге могут искажать реальную картину функционирования инфокоммуникационных систем/сетей.

Применение регрессионного анализа в задачах теории телетрафика особенно интересно. Часто встречающиеся задачи являются задачами оптимизации. И здесь применение метода наименьших квадратов более, чем уместно, поскольку по своей сути этот метод направлен на определении параметров модели тренда, которая описывает тенденции развития во времени случайного явления/процесса. Таким образом, появляется возможность прогнозирования поведения инфокоммуникационной системы и выбора оптимальной модели ее функционирования.

Литература

1. *R. Hyndman, G. Athanasopoulos*. (2016). Forecasting: Principles and Practice. 2td ed. Melbourne, Australia. 291p. (in English)
2. *H. Wickham*. (216). Elegant graphics for data analysis. 2td ed. Springer. 213p. (in English)
3. *G. Athanasopoulos, R.J. Hyndman, N. Kourentzes, F. Petropoulos*. (2017). Forecasting with temporal hierarchies. European Journal of Operational Research, *262*(1), pp.60–74. (in English)
4. *C. Bergmeir, R.J. Hyndman, J.M. Benítez*. (2016). Bagging exponential smoothing methods using STL decomposition and Box-Cox transformation. International Journal of Forecasting, *32*(2), pp.303–312. (in English)
5. *C. Bergmeir, R.J. Hyndman, B. Koo*. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics and Data Analysis, *120*, pp.70–83. (in English)
6. *S.N. Lahiri*. (2003). Resampling methods for dependent data. New York, USA: Springer Science & Business Media. 374p. (in English)
7. *J.K. Ord, R. Fildes, N. Kourentzes*. (2017). Principles of business forecasting. 2td ed. Wessex Press Publishing Co. (in English)
8. *S.L. Wickramasuriya, G. Athanasopoulos*. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. J American Statistical Association, 114(526), pp. 804–819. (in English)
9. *F.E. Harrell*. (2015). Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. New York, USA: Springer. 568p. (in English)
10. *K. Madsen, H.B. Nielsen, O. Tingleff*. (2004). Methods for Non-linear Least Squares Problem Cobenhavn, Technical University of Denmark, 2004. 30p.
11. *Alfred DeMaris*. (2004). Regression with Social Data, Modeling Continuous and Limited Response Variables 2004, John Wiley & Sons, Inc; 558p.
12. *G.G. Vining, S. Kowalski*. (2010). Statistical Methods for Engineers Duxbury Press, 2010. 648p.
13. *M.H. Kutner, C.J. Nachtsheim, J. Neter, W. Li*. (2004). Applied Linear Statistical Models McGraw-Hill, 2004.1424 p.
14. *M.H. DeGroot, M.J. Schervish*. (2011). Probability and Statistics Addison Wesley, 4th Edition. 2011. 911p.

REGRESSION ANALYSIS APPLICATION FOR TELETRAFFIC THEORY TASKS

Alina A. Sherstneva, *SibSUTIS, Novosibirsk, Russia,* ***asherstneva@sibguti.ru***

Abstract

The article aims to consider least squares approach for solving problems of queuing systems theory. The opportunity of predicting the behavior of infocommunication system is shown. Choosing the optimal model of its functioning is proposed. On base monitoring system metrics, statistical data were formed. The article proposes to make data trend forecasting, to estimate parameters of random processes over time. To obtain the results of functioning data in infocommunication systems that are as close as possible to the real values, polynomial and sine models are considered. The method of regression analysis is proposed to determine the parameter values for a model from a set of observational data. In theoretical research, the linear and nonlinear least squares methods are used in terms of a circle. The task of experimental analysis is to obtain an estimated parameter of sine, polynomial models and the center of circle. Experimental analysis was performed using the mathematical modeling program Matlab. A uniformly distributed random sequence and a random sequence with normal distribution are generated. The sequence with experimental data for polynomial and sine models, respectively, are calculated. The correspondence each model for generated data is shown in graphical form. The measurement data obeys observations. The estimated parameters are summarized in the tables. The polynomial order is estimated. The estimated dispersion curve of the polynomial model is obtained. The calculated variance values of the polynomial model are presented. Data trend forecasting for measurement data is made. The estimated values are extremally close to real data. The results are shown in graphs. Finally, an approximate model of the circumference of measurement data is presented in graphical form. After some iterations with estimated center from the arithmetic mean the new circle center is given. And quite close values for center and radius of circle are obtained.

Keywords: *regression analysis, least squares approach, sine model, polynomial model, forecasting, data trend, estimation, predictor variables, machine learning, statistical metrics, observation; measurement, data mining.*

References

1. R. Hyndman, G. Athanasopoulos. (2016). Forecasting: Principles and Practice. 2td ed. Melbourne, Australia. 291p. (in English)
2. H. Wickham. (216). Elegant graphics for data analysis. 2td ed. Springer. 213p. (in English)
3. G. Athanasopoulos, R.J. Hyndman, N. Kourentzes, F. Petropoulos. (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research,* 262(1), pp.60-74. (in English)
4. C. Bergmeir, R.J. Hyndman, J.M. Benetez. (2016). Bagging exponential smoothing methods using STL decomposition and Box-Cox transformation. *International Journal of Forecasting*, 32(2), pp.303-312. (in English)
5. C. Bergmeir, R.J. Hyndman, B. Koo. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics and Data Analysis,* 120, pp.70-83. (in English)
6. S.N. Lahiri. (2003). Resampling methods for dependent data. New York, USA: Springer Science & Business Media. 374p. (in English)
7. J.K. Ord, R. Fildes, N. Kourentzes. (2017). Principles of business forecasting. 2td ed. Wessex Press Publishing Co. (in English)
8. S.L. Wickramasuriya, G. Athanasopoulos. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *J American Statistical Association*, 114(526), pp. 804-819. (in English)
9. F.E. Harrell. (2015). Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. New York, USA: Springer. 568p. (in English)
10. K. Madsen, H.B. Nielsen, O. Tingleff. (2004). Methods for Non-linear Least Squares Problem Cobenhavn, Technical University of Denmark, 2004. 30p.
11. Alfred DeMaris. (2004). Regression with Social Data, Modeling Continuous and Limited Response Variables 2004, John Wiley & Sons, Inc; 558p.
12. G.G. Vining, S. Kowalski. (2010). Statistical Methods for Engineers Duxbury Press, 2010. 648 p.
13. M.H. Kutner, C.J. Nachtsheim, J. Neter, W. Li. (2004). Applied Linear Statistical Models McGraw-Hill. 1424 p.
14. M.H. DeGroot, M.J. Schervish. (2011). Probability and Statistics Addison Wesley, 4th Edition. 911 p.

Information about author:

Alina A. Sherstneva, *Candidate of Tech. Sciences, associated professor, Siberian State University of Telecommunications and Information Sciences, Department of Electrical Communication, Novosibirsk, Russia*