Лекция 13. Решение задачи регрессии в рамках статистического и детерминистского подходов

Задача регрессии является одной из классических задач анализа данных и относится к классу предсказательных (predictive) задач.

Предсказательные задачи осуществляются в два этапа. **Первый этап** направлен на построение модели анализируемого объекта (процесса) путем обработки данных – получаемых в ходе экспериментов наблюдений, в той или иной степени, его характеризующих. **На втором этапе** модель используется для предсказания результатов по отношению к новым наборам данных.

Специфика задачи регрессии состоит в том, что модель объекта строится в виде функциональной зависимости набора непрерывных выходных переменных от заданного набора входных, а второй этап направлен на определение значений выходных переменных при появлении новых значений входных. Общность задач оценивания и регрессии определяется тем, что построение регрессионной модели фактически сводится к задаче оценивания неизвестных параметров модели объекта с использованием схожих методов и алгоритмов.

Итак, задача регрессии состоит в установлении функциональной зависимости между зависимыми и независимыми показателями и переменными в понятной форме. Обычно вектор входных (независимых) переменных (называемых также регрессорами или предикторами) $x \in X \subseteq \boldsymbol{R}^n$, а вектор выходных (зависимых) переменных $y \in Y \subseteq \boldsymbol{R}^1$.

В случае, если $n=1$, говорят о постановке задачи парной регрессии, а если $n>1$, – множественной регрессии. В более широкой постановке может рассматриваться несколько зависимых переменных, которые объединяются в вектор $y \in R^h$.

Если значение зависимой переменной определяется по отношению к будущему моменту времени, то задача регрессии называется **задачей прогнозирования**.

При решении задачи требуется на основе имеющихся исходных данных установить функциональную зависимость $y = \tilde{\varphi}(x)$, которая с приближением отражает реальную зависимость $y = \varphi(x)$ и которую можно далее использовать для предсказания неизвестного значения зависимой переменной для любых значений независимых, входных переменных). Приближение должно удовлетворять некоторому критерию качества.

1. Постановка и решение задачи регрессии в рамках статистического подхода

Задача может решаться как в статистической, так и в детерминистской постановке.

В первом случае x и y рассматриваются как значения случайных величин или случайных векторов. При этом считается, что известны полные статистические описания этих величин (векторов) в виде совместной плотности распределения вероятностей $p(x, y)$.

На практике чаще все-таки возникает ситуация, когда статистические описания переменных неизвестны или их не интерпретируют как случайные величины. В то же время, имеются обучающие наборы измеряемых (наблюдаемых) данных, характеризующие реализации (выборочные значения) этих величин.

При этом точная функциональная зависимость не может быть получена, поскольку эти величины либо являются СВ, между которыми существует статистическая связь и на измерения накладываются случайные ошибки, либо между этими величинами существует функциональная связь, но наблюдения не в полной мере достоверны из-за ошибок измерения.

Тогда задача регрессии решается в рамках детерминистского подхода на основе метода наименьших квадратов.

1. Постановка и решение задачи регрессии в рамках детерминистского подхода по методу наименьших квадратов

Пусть дана обучающая выборка данных $X^N = \{x^{(1)},...,x^{(N)}\}$, $x^{(i)} \in X \subseteq \boldsymbol{R}^n$, в которой каждому значению вектора-столбца независимых (входных) переменных соответствует совокупность значений зависимой (выходной) переменной $Y^N = \{y^{(1)},...,y^{(N)}\}$, $y^{(i)} \in Y \subseteq \boldsymbol{R}^1$ определены на множестве континуум. Требуется построить отображение вида $\tilde{\varphi}: X \to Y$, т.е. алгоритм восстановления регрессии, аппроксимирующие реально действующую зависимость переменных с заданным критерием качества.

Изначально всегда вводится ограничение на вид используемых для аппроксимации функций, которые задаются в виде семейства $\psi(x,a)$, $a \in R^m$, где a – вектор неопределенных параметров модели. В рамках метода наименьших квадратов требуется найти значение $\tilde{a}$, для которого минимизируется функционал

$$J(a) = \sum_{i=1}^{N} (\psi(x^{(i)},a) - y^{(i)})^2 \to \min_a, \tag{1}$$

Для (4) как необходимое условие минимума выполняется

$$\left.\frac{\partial J}{\partial a}\right|_{a=\tilde{a}} = 2\sum_{i=1}^{N} (\psi(x^{(i)},a) - y^{(i)}) \frac{\partial \psi(x^{(i)},a)}{\partial a} = 0. \tag{2}$$

Полученная система из m уравнений имеет численное решение, которое позволяет построить регрессионную зависимость в виде

$$\tilde{y} = \psi(x,\tilde{a}). \tag{3}$$

Линейная параметрическая регрессия. Рассмотрим первоначально решение задачи построения линейной регрессии на основе (1-3). Тогда в качестве $\psi(x,a)$ используется линейная функция

$$\psi(x,a) = a_0 + \sum_{k=1}^{n} a_k x_k ,$$

где a_0 – скалярная величина; $a_k = (a_1,...,a_n)^T$ – вектор коэффициентов при компонентах входного вектора размера $n \times 1$. Введем расширение вектора x, добавляя фиктивную компоненту $x_0 \equiv 1$, $x_e = (x_0, x_1,...,x_n)^T$ и объединяют коэффициенты в общий вектор $a = (a_0, a_1,...,a_n)^T$. Тогда

$$\psi(x,a) = a^T x_e = \sum_{k=0}^{n} a_i x_i .$$

Введем матрицу X размера $N \times (n+1)$, строками которой являются транспонированные расширенные векторы входной обучающей выборки, и вектор выходных данных обучающей выборки

$$X = \begin{pmatrix} x_e^{(1),T} \\ x_e^{(2),T} \\ \vdots \\ x_e^{(N),T} \end{pmatrix} = \begin{pmatrix} 1 & x_1^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^{(N)} & \cdots & x_n^{(N)} \end{pmatrix}, \; y = (y^{(1)},..., y^{(N)})^T = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{pmatrix}.$$

Из общего решения (2) получим следующую систему $n+1$ уравнений в матричном виде, обеспечивающих минимизацию невязки $Xa - y$

$$J(a) = \|Xa - y\|^2 = (Xa - y)^T (Xa - y) \to \min_a$$

$$\left.\frac{\partial J}{\partial a}\right|_{a=\tilde{a}} = 2X^T(Xa - y) = 2\sum_{i=1}^{N}(a^T x_e^{(i)} - y^{(i)})x_e^{(i)} = 0, \;\; \frac{\partial \psi(x^{(i)}, a\,)}{\partial a} = \frac{\partial a^T x_e^{(i)}}{\partial a} = x_e^{(i)}$$

Тогда исходная система уравнений для нахождения $\tilde{a}$ перепишется в виде другой системы, называемой нормальной системой, которая в матричной форме имеет следующий вид:

$$X^T y - X^T X\tilde{a} = 0.$$

Нормальная система всегда совместна. Матрица $H = X^T X$ является квадратной матрицей размера $(n+1)\times(n+1)$.

В случае, если матрица $H = X^T X$ является невырожденной, существует единственное, так называемое нормальной решение

$$\tilde{a} = (X^T X)^{-1} X^T y = X^+ y,\ X^+ = (X^T X)^{-1} X^T. \qquad (4)$$

Матрица $X^+ = (X^T X)^{-1} X^T$ является в этом случае псевдобратной матрицей для X. Она обладает свойствами, во многом схожими со свойствами обратной матрицы, а именно

$$X^+ X = (X^T X)^{-1} X^T X = I, \qquad X^+ X X^+ = (X^T X)^{-1} X^T X (X^T X)^{-1} X^T = X^+, \qquad X X^+ X = X (X^T X)^{-1} X^T X = X.$$

Отметим, что если матрица X является квадратной (объем обучающей выборки равен числу оцениваемых параметров регрессии) и невырожденной, то полученное решение является решением системы линейных алгебраических уравнений (СЛАУ) $Xa = y$ по правилу Крамера: $\tilde{a} = X^{-1} y$. При этом невязка тождественно равна нулю.

Однако на практике, в основном, СЛАУ $Xa = y$, является переопределенной (объем обучающей выборки превышает число оцениваемых параметров регрессии). В этом случае вектор $\tilde{a}$ может быть определен на основе (4). Таким образом, окончательное решение задачи имеет вид

$$\tilde{y} = \psi(x, \tilde{a}) = \tilde{a}^T x = \tilde{a}_0 + \sum_{k=1}^{n} \tilde{a}_i x_i. \qquad (5)$$

Величины

$$e^{(i)} = y^{(i)} - \tilde{y}^{(i)}, i = \overline{1,N} \ , \quad \tilde{y}^{(i)} = \tilde{a}^T x^{(i)}$$

в (8) называются остатками. Анализ остатков весьма важен для анализа полученного решения с точки зрения его адекватности и пригодности полученной модели к использованию. При реализации МНК сумма остатков равна нулю

$$\sum_{i=1}^{N} e^{(i)} = \sum_{i=1}^{N} (y^{(i)} - \tilde{y}^{(i)}) = 0 .$$

В качестве самостоятельного задания можно предложить рассмотреть решение задачи линейной регрессии для случая, когда выходная реакция является вектором $y^{(i)} \in Y \subseteq R^h$.

Проверка значимости модели регрессии. Анализ полученной математической модели регрессии предполагает использование различных методов и алгоритмов математической статистики.

Одной из центральных задачей анализа уравнения регрессии является проверка адекватности и значимости выбранной модели.

Проверить адекватность уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным, и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной. Это нужно сделать для того, чтобы убедиться в пригодности и надежности модели.

Проверка адекватности регрессионной модели основана на сравнении рассеивания экспериментальных значений наблюдаемой переменной относительно линии регрессии с рассеиванием этих значений относительно ее среднего значения $\bar{y}$, которое с учетом свойств суммы остатков равно

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y^{(i)} = \frac{1}{N}\sum_{i=1}^{N} \tilde{y}^{(i)} = \frac{1}{N}\sum_{i=1}^{N} \tilde{a}^T x^{(i)} = \tilde{a}^T \bar{x}, \quad \tilde{y}^{(i)} = \tilde{a}^T x^{(i)}.$$

В проверке модели используется разложение общей суммы квадратов отклонений переменной y от среднего, которое имеет вид

$$TSS = \sum_{i=1}^{N} (y^{(i)} - \bar{y})^2 = \sum_{i=1}^{N} (\tilde{y}^{(i)} - \bar{y})^2 + \sum_{i=1}^{N} (y^{(i)} - \tilde{y}^{(i)})^2 = ESS + RSS,$$

Левая часть равенства – это общая сумма квадратов отклонений (total sum of squared) –TSS, которая равна сумме квадратов отклонений от среднего, объясненных регрессией (explained sum of squared) – ESS и остаточной сумме квадратов (residual sum of squared) – RSS, определяемая случайными остатками.

Чем меньше RSS, тем регрессия лучше. Чем меньше ESS, тем регрессия хуже.

Несмещенной оценкой дисперсии ошибки измерения выходной переменной или случайной ошибки шума, действующего при получении наблюдений, является величина остаточной дисперсии

$$\sigma_{\tilde{y}}^2 = \frac{RSS}{N-n-1} = \sum_{i=1}^{N} (y^{(i)} - \tilde{y}^{(i)})^2.$$

где $N-n-1=m$ – число степеней свободы, которое определяется как разность между числом точек, в которых ставились эксперименты N, по которым оценивают коэффициенты a и числом этих коэффициентов $n+1$.

Для проверки адекватности уравнения регрессии необходимо, чтобы число степеней свободы было больше нуля.

При отрицательном числе степеней свободы, т. е. при ($N < n+1$) метод наименьших квадратов не может быть использован.

При $m = 0$ уравнение регрессии может быть получено, однако статистический анализ этого уравнения провести нельзя, ибо для проверки адекватности модели не остается степеней свободы.

При $m > 0$ остаются степени свободы для проверки адекватности регрессионной модели. Вводится **коэффициент детерминации**

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

Этот коэффициент показывает качество подгонки регрессионной модели к наблюдаемым значениям.

Если он равен нулю, то регрессия на y на x не улучшает качества предсказания значения y' по сравнению с тривиальным предсказанием $y' = \bar{y}$.

Если он равен своему максимальному значению – единице, то это означает точную подгонку регрессии под экспериментальные данные (все остатки равны нулю и имеет место строгая функциональная зависимость).

Для приемлемых моделей предполагается, что коэффициент детерминации должен быть не меньше 50 %.

Иногда коэффициент детерминации дает слишком оптимистические результаты, поскольку значение R^2 возрастает с ростом числа переменных (регрессоров), что не означает улучшения качества предсказания, и потому вводится **скорректированный коэффициент детерминации**

$$R^2 = 1 - \frac{RSS/(N-n-1)}{TSS/(N-1)}.$$

При реализации МНК используются также методики проверки гипотезы о нулевых значениях различных комбинаций коэффициентов регрессии.

Проверка значимости всей модели предполагает проверку гипотезы H_0 о равенстве нулю всех коэффициентов регрессии, кроме коэффициента a_0 при фиктивной переменной $x_0 \equiv 1$. В этом случае рассматривается отношение оценок дисперсий

$$F = \frac{ESS/n}{RSS/(N-n-1)}.$$

Данная величина при несмещенном, гауссовском характере случайных ошибок измерения выходной переменной имеет статистику распределения Фишера $F(n, N-n-1)$. В общем случае произвольного распределения ошибок такая статистика является асимптотической. Если

$$F < F_\gamma(n, N-n-1),$$

то гипотеза H_0 о равенстве нулю всех коэффициентов, кроме, может быть, a_0, подтверждается. Здесь F_γ – квантиль F - распределения уровня доверительной вероятности $1-\gamma$, показывающий пороговый уровень статистики, который превышается с вероятностью γ. Это означает, что необходимо исключать часть входных переменных из рассмотрения.

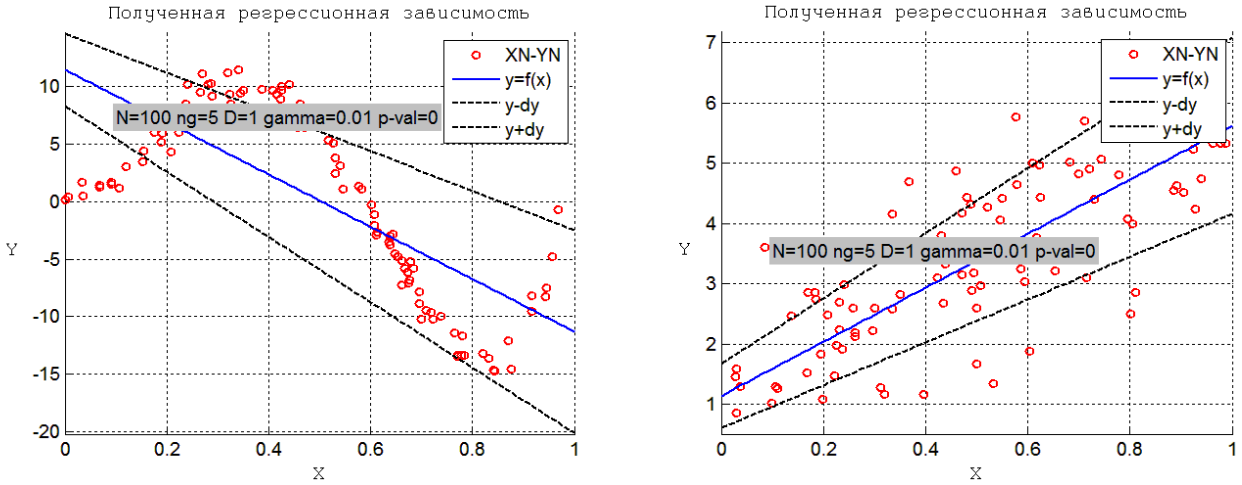
Следует отметить, что при применении критерия Фишера может использоваться и обратное отношение остаточной дисперсии к дисперсии отклонений регрессии от среднего. Тогда используют статистику $F(N-n-1,n)$. Обычно тест применяют так, чтобы в числителе находилось потенциально большая величина.

В этом случае удобным способом проверки гипотез является использования p-значения – вероятности того, что случайная величина с данным распределением Фишера превысит данное значение статистики. Если p-значение меньше уровня значимости, то нулевая гипотеза отвергается, в противном случае принимается.

Далее будут рассмотрены несколько примеров, иллюстрирующих возможности программной реализации методов регрессионного анализа. Для этого использовались возможности среды MATLAB (раздел Statistics Toolbox), содержащей большое количество стандартных функций для построения регрессий различного вида, анализа регрессионных моделей и обеспечения возможностей построения и визуализации регрессионных характеристик в интерактивном режиме.

В представленных ниже примерах реализуется построение регрессионной линейной модели для различных вариантов задания исходной восстанавливаемой модели (в виде полинома, в виде гармонического ряда). Проводится генерация обучающей выборки данных $X^N=\{x^{(1)},...,x^{(N)}\}$, $x^{(i)}\in X\subseteq \boldsymbol{R}^n$, как реализаций входного вектора x размерности $n=1$ с равномерным распределением в заданном диапазоне значений. Значения выходной переменной $Y^N=\{y^{(1)},...,y^{(N)}\}$, $y^{(i)}\in Y\subseteq R^1$ генерируются как $y^{(i)}=a^T x^{(i)}+\alpha_i$, где α_i, $i=1,N$ центрированные значения гауссовской ошибки с дисперсией $D=\sigma_\alpha^2$.

Кроме того, в примерах проводится вычисление коэффициента детерминации R^2, статистики Фишера $F(n, N-n-1)$, расчет значения F-статистики, который превышается с вероятностью gamma, расчет вероятности превышения полученного значения F (p-значение). Полученные значения этих величин, показывают, что полученная регрессионная модель является статистически значимой (p_value не превосходит gamma). На графиках рис.1а,б показаны зависимости для различных вариантов задания восстанавливаемой модели.



а) б)

Рис.1. Примеры восстановления регрессии: исходная зависимость линейная (var1=1) – а; исходная зависимость в виде гармонического ряда (var1=2) – б

Как следует из представленных графиков, использование линейной регрессии возможно и в случае, если исходная модель является существенно нелинейной. Однако, если изначально известен характер зависимости, для этого целесообразно использовать другие варианты алгоритма восстановления регрессии на основе МНК, которые будут рассмотрены ниже.

Расширения линейной регрессии и нелинейная регрессия. Очевидно, что полученное выше решение задачи регрессии может быть использовано при произвольном нелинейном преобразовании $x : \bar{x} = \psi(x)$, если неизвестные параметры входят в (1-3) линейным образом:

$$\psi(x,a) = a^T \psi(x), \quad \tilde{y} = \tilde{a}^T \psi(x).$$

Пусть отображение $\psi : R^n \to R^{s+1}$ осуществляет перевод исходного пространства входных переменных в новое пространство в общем случае другой размерности. Тогда $\psi(x) = (1, \psi_1(x), ..., \psi_s(x))^T$. Вводя для всех элементов обучающей выборки замену $\bar{x}^{(i)} = \psi(x^{(i)}),\ i = \overline{1,N}, \bar{x} \in \bar{X} \subseteq \boldsymbol{R}^s$, можно перейти к поиску решения задачи линейной регрессии для новых переменных

$$\tilde{y} = \psi(x, \tilde{a}) = \tilde{a}_0 + \sum_{k=1}^{s} \tilde{a}_i \bar{x}_i = \tilde{a}^T \bar{x}_e,$$

где $\tilde{a}$ ищется на основе соотношений, аналогичных (4)-(6).

В качестве примера можно привести использование представление полиномов n переменных различной степени до d включительно в виде линейной функции $s+1$ переменных $s+1 = C_{n+d}^n$ (с добавлением фиктивной переменной $\bar{x}_0 \equiv 1$)

$$\bar{x}_0 = 1,\ \bar{x}_i = x_i,\ i = \overline{1,n},\quad \bar{x}_{n+1} = x_1 x_1,\ \bar{x}_{n+2} = x_1 x_2, ...,\ \bar{x}_{n(n+1)-1} = x_n x_{n-1},\quad \bar{x}_{n(n+1)} = x_n x_n, ...$$

При таком представлении каждое слагаемое искомой регрессии имеет вид

$$a_k \psi_k(x) = a_{i_1 \ldots i_n} x_1^{i_1} \times x_2^{i_2} \times \ldots x_n^{i_n}, i_1 + i_2 + \ldots i_n \le d, \quad k = \overline{1, s+1},$$

где $a_{i_1 \ldots i_n}$ – исходные коэффициенты, которые приравниваются к коэффициентам линейной регрессии a_k после замены переменных. В результате всегда можно представить исходный полином, описывающий линейную регрессию по отношению к s новым входным переменным. Следует еще раз обратить внимание, что при такой замене элементы исходной обучающей выборки преобразуются в элементы обучающей выборки для новых переменных эквивалентным образом.

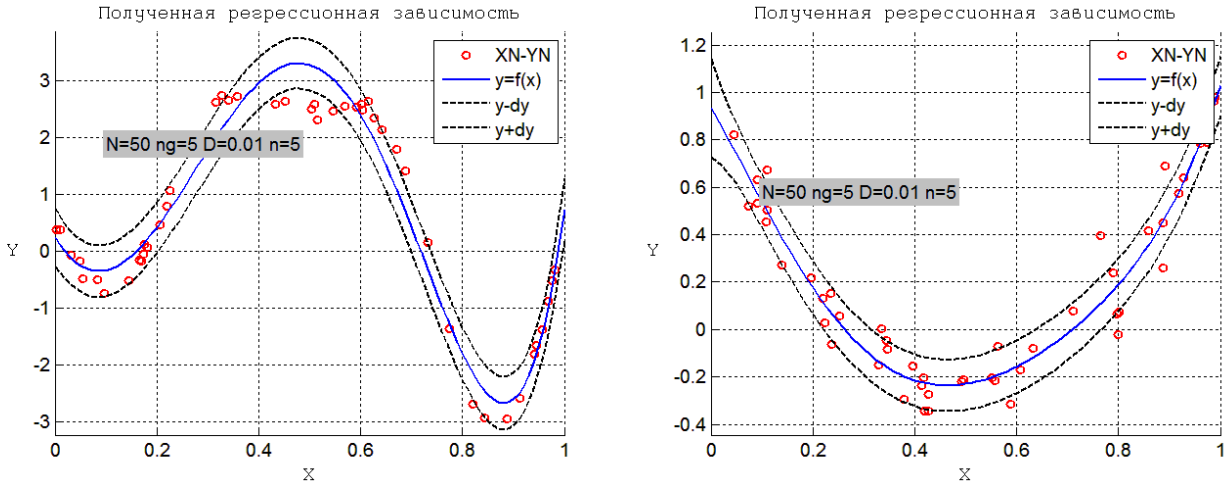
В частном случае, когда входная переменная является скалярной величиной, регрессия ищется изначально в виде

$$\psi(x, a) = a^T \psi(x) = \sum_{k=0}^{d} a_k x^k .$$

Такая регрессия в литературе называется полиномиальной и ее использование позволяет повысить точность представления данных в случаях, когда исходная модель, определяющая вид восстанавливаемой зависимости, является нелинейной.

В представленной ниже примере реализуется построение полиномиальной регрессии модели для различных вариантов задания исходной, восстанавливаемой модели (в виде полинома, в виде гармонического ряда). Проводится генерация обучающей выборки данных $X^N = \{x^{(1)}, \ldots, x^{(N)}\}$, $x^{(i)} \in X \subseteq R^n$, как реализаций входного вектора x размерности $n = 1$ с равномерным распределением в заданном диапазоне значений. Значения выходной переменной $Y^N = \{y^{(1)}, \ldots, y^{(N)}\}$, $y^{(i)} \in Y \subseteq R^1$ генерируются как $y^{(i)} = a^T x^{(i)} + \alpha_i$, где α_i, $i = 1, N$ центрированные значения гауссовской ошибки с дисперсией $D = \sigma_\alpha^2$.

На рис.2а,б представлены полученные зависимости построенной полиномиальной регрессии и границ доверительных интервалов для двух случаев: когда в качестве восстанавливаемой исходной модели используется полиномиальная модель, и, когда в качестве восстанавливаемой исходной модели используется модель в виде гармонического ряда.



а) б)

Рис.2. Примеры восстановления полиномиальной регрессии: исходная полиномиальная зависимость – а; исходная зависимость в виде гармонического ряда – б

Анализ представленных графиков показывает, что в данном случае использование регрессии полиномиального типа позволяет достаточно эффективно восстанавливать нелинейные зависимости различного вида.

Таким образом, можно видеть, что использование линейной регрессии после нелинейного преобразования исходных входных переменных в большинстве случаев позволяет решить задачу восстановления нелинейной функциональной зависимости произвольного вида.