Лабораторная работа 4

Построение уравнения линейной регрессии по результатам измерений

Пусть требуется исследовать зависимость $y = y(x)$, причем величины y и x измеряются в одних и тех же экспериментах. Без ограничения общности можно считать, что величина x измеряется точно, в то время как измерение величины y содержит случайные погрешности. Это означает, что погрешность измерения величины x пренебрежимо мала по сравнению с погрешностью измерения величины y.

Таким образом, результаты эксперимента можно рассматривать как выборочные значения случайной величины $\eta(x)$, зависящей от x как от параметра. Регрессией называют зависимость условного математического ожидания этой величины от x, т.е. $y(x) \equiv \mathbf{M}(\eta|x)$. Задача регрессионного анализа состоит в восстановлении функциональной зависимости $y(x)$ по результатам измерений

$$(x_i, y_i), i = 1,2, \dots, n.$$

Аппроксимируем искомую зависимость $y(x)$ функцией $\mathbf{f}(\mathbf{x}, a_1, a_2, \dots, a_k)$. Это означает, что результаты измерений можно представить в виде

$$y_i = \mathbf{f}(\mathbf{x}, a_1, a_2, \dots, a_k) + \xi_i,$$

где $a_1, a_2, \dots, a_k -$ неизвестные параметры регрессии; $\xi_i -$ случайные величины, характеризующие погрешности эксперимента.

Обычно предполагается, что $\xi_i -$ это независимые нормально распределенные случайные величины с $\mathbf{M}(\xi_i) = 0$ и одинаковыми дисперсиями $\mathbf{M}\left({\xi_i}^2\right) = \sigma^2$.

Параметры $a_1, a_2, \dots, a_k$ следует выбрать такими, чтобы отклонение предложенной функциональной зависимости от результатов эксперимента было минимальным.

Часто в качестве меры отклонения принимают величину

$$\Phi(a_1, a_2, \dots, a_k) = \sum_{i=1}^{n} (f(x_i, a_1, a_2, \dots, a_k) - y_i)^2$$

и, следовательно, параметры $a_1, a_2, \dots, a_k$ определяются методом наименьших квадратов.

На практике регрессионный анализ состоит из трёх этапов. На первом этапе выдвигают гипотезу о виде функции $\mathbf{f}(\mathbf{x}, a_1, a_2, \dots, a_k)$, на втором – по имеющимся данным находят оценки неизвестных параметров $a_1, a_2, \dots, a_k$. На третьем этапе проверяют согласие выдвинутой гипотезы с результатами измерений.

Рассмотрим простейший случай, а именно линейную регрессию. Пусть выдвинута гипотеза о том, что функция f имеет вид

$$f(x, a_0, a_1) = a_0 + a_1 \cdot x.$$

Найдём оценки параметров a_0 и a_1 методом наименьших квадратов. Для этого минимизируем функцию

$$\Phi(a_0, a_1) = \sum_{i=1}^{n} \big((a_0 + a_1 \cdot x) - y_i\big)^2.$$

Приравнивая нулю производные $\frac{\partial \Phi}{\partial a_0}$ и $\frac{\partial \Phi}{\partial a_1}$, получаем

$$a_0 = \frac{\sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$a_1 = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}.$$

Проверяя согласие построенной линии регрессии с результатами эксперимента, можно руководствоваться следующими соображениями. Идея любой регрессии состоит в том, чтобы часть изменений измеряемой величины

связать с изменением внешних переменных (в рассматриваемом случае только одна внешняя переменная x). Не предполагая, что y зависит от x, можно было бы за меру разброса результатов эксперимента принять величину $\varepsilon_1 = \sum_{i=1}^{n}(y_i - \bar{y})^2$, где $\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i$. Если прямая регрессии построена, то за меру разброса естественно принять сумму квадратов отклонений от линии регрессии, т.е. величину

$$\varepsilon_2 = \sum_{i=1}^{n}(y_i - a_0 - a_1 \cdot x_i)^2.$$

Если $\varepsilon_2 \approx \varepsilon_1$, то это значит, что аппроксимирующая функция выбрана неудачно, т.е. подходящую функцию регрессии следует искать не среди прямых, а, например, среди парабол или кривых другого вида.

Задание:

1. Составить m-функцию, содержащую:

1.1 ввод из файла (текстовый файл с именем data.txt) и печать числового массива, содержащего n измеренных значений

$$(x_i, y_i), i = 1,2, \dots, n;$$

1.2 вычисление коэффициентов линейной регрессии a_0 и a_1;

1.3 вычисление и печать меры разброса ε_1 и меры отклонения от регрессии ε_2.

2. Построить график, содержащий исходные экспериментальные данные и кривую линейной регрессии.

3. Сравнить величины ε_1 и ε_2 и сделать выводы о пригодности или непригодности линейной регрессии в качестве аппроксимирующей функции.

Варианты для лабораторной работы по аппроксимации

Значения $x_i = i \cdot 0{,}1, i = 1,2, \dots, 20$, одинаковы для всех вариантов

Варианты				
1	2	3	4	5
Значения $y_i = y(x_i)$				
5,998	6,030	5,850	6,310	5,650
5,820	6,072	5,619	6,308	5,431
5,754	6,297	5,569	6,546	5,250
5,828	6,428	5,426	6,855	5,000
5,627	6,425	5,237	7,073	4,790
5,597	6,473	5,025	7,770	4,569
5,693	6,592	4,988	7,225	4,296
5,469	6,815	5,037	7,739	4,065
5,413	6,786	4,586	7,995	3,837
5,526	6,925	4,575	8,963	3,519
5,344	7,116	4,445	8,247	3,281
5,304	7,053	4,353	8,472	2,926
5,352	7,224	3,933	8,627	2,801
5,301	7,439	3,899	8,936	2,546
5,424	7,302	3,793	9,082	2,232
4,996	7,426	3,473	9,976	2,016
5,080	7,797	3,551	9,363	1,794
5,256	7,871	3,171	9,679	1,663
5,090	7,929	3,330	9,846	1,375
5,053	8,060	3,044	10,013	1,217

Варианты				
6	7	8	9	10
Значения $y_i = y(x_i)$				
6,323	3,8812	4,0823	3,9023	4,0302
6,523	3,8604	4,1842	3,8310	4,2339
6,646	3,8401	4,3803	3,6028	4,4903
7,256	3,9123	4,4614	3,4733	4,7195
7,487	3,7139	4,4420	3,3103	5,0010
7,827	3,4930	4,5522	3,0502	5,2605
8,133	3,5108	4,6612	3,1415	5,3633
8,402	3,6833	4,8901	2,8318	5,8711
8,581	3,7408	4,8632	2,6611	5,6703
9,014	3,4712	5,0410	2,5302	5,8911
9,049	3,6032	5,2214	2,3503	6,1601

9,571	3,5122	4,9010	2,4910	6,6505
9,891	3,4810	5,3907	2,1901	6,3902
10,073	3,3043	5,5633	1,8232	6,8104
10,406	3,2342	5,4214	1,6973	7,0810
10,821	3,2600	5,8521	1,5403	7,2410
11,151	3,1411	5,9903	1,2223	7,6133
11,232	3,1717	5,8541	1,1793	7,6436
11,655	2,9603	6,0113	1,0433	8,0393
11,952	2,8193	5,9739	1,1203	7,9243

Варианты				
11	12	13	14	15
Значения $y_i = y(x_i)$				
3,8220	4,2710	1,9210	2,1401	1,5636
3,4432	4,4523	1,9123	2,1923	1,8440
3,1692	4,8416	2,0932	2,3203	1,5110
2,9581	5,1413	1,7317	2,5932	1,5233
2,7335	5,5503	1,8803	2,5603	1,0903
2,4010	5,8500	1,8113	2,6401	1,0411
2,2732	6,1893	1,7173	2,6633	1,0561
1,8536	6,3820	1,6603	2,8405	0,9122
1,8830	6,7210	1,4738	3,0403	0,6950
1,3230	7,0402	1,4434	2,9405	0,5111
1,1824	7,2632	1,2322	3,2303	0,4643
1,1508	7,7023	1,3732	3,2773	0,1422
0,8554	7,7803	1,3020	3,3125	-0,0610
0,4833	8,3312	1,2231	3,1300	-02929
0,1821	8,6203	1,3816	3,4926	-0,2831
-0,0113	8,7825	1,3523	3,5603	-0,2524
-0,1214	9,0602	1,1401	3,6604	-0,5732
-0,6030	9,5642	1,0024	3,7904	-0,5763
-0,6836	9,7108	0,9693	3,9636	-1,0603
0,5436	10,1403	0,9329	4,0810	-1,0122