УДК 62-50

ПАВЛОВ А.А.,
КАЛАШНИК В.В.

РЕКОМЕНДАЦИИ ПО ВЫБОРУ ЗОНЫ ПРОВЕДЕНИЯ АКТИВНОГО ЭКСПЕРИМЕНТА ДЛЯ ОДНОМЕРНОГО ПОЛИНОМИАЛЬНОГО РЕГРЕССИОННОГО АНАЛИЗА

В статье рассматривается выбор зоны проведения активного эксперимента так, чтобы уменьшить дисперсии оценок коэффициентов регрессионного полинома. Проводится сравнение различных промежутков, и приводятся таблицы с полученными дисперсиями оценок коэффициентов на них. Рассматривается возможность масштабирования уже полученных исходных данных.

The choice of the area of an active experiment execution is considered so that the dispersions of the estimates of regression polynomial's coefficients be decreased. The comparison of different gaps is fulfilled and the tables with obtained dispersions of the coefficients' estimates for them are given. A possibility of zooming already retrieved data is considered.

1. Вступление

Проблема построения нелинейной регрессии по данным с шумом является одной из наиболее востребованных на практике задач. Значительную роль играет соответствие построенной регрессионной модели реальной зависимости. В данной статье приводятся рекомендации, относительно зоны проведения активного эксперимента, над результатами которого будет выполняться регрессионный анализ. Это касается только тех случаев, когда исследователь может управлять проведением эксперимента, а именно он имеет возможность выбирать диапазон аргумента. Рассматривается одномерный полиномиальный регрессионный анализ с использованием нормированных ортогональных полиномов Форсайта, приведенных в главе 6 [1]. Критерием улучшения регрессионной модели является уменьшение дисперсий оценок коэффициентов в регрессионном полиноме. В статье приводится аналитическое доказательство того, что при масштабировании интервала, после проведения эксперимента, не повлияет на качество регрессии.

2. Выбор зоны проведения активного эксперимента

Можно выделить шесть условных зон проведения эксперимента:

1) Начало больше нуля, малый интервал. Например, [10;20].
2) Начало больше нуля, интервал большой. Например, [10;1010].
3) Начало в точке ноль, интервал малый. Например, [0;10].
4) Начало в точке ноль, интервал большой. Например, [0;1000].
5) Симметричен относительно точки ноль, интервал мал. Например, [-5;5].
6) Симметричен относительно точки ноль, интервал большой. Например, [-500;500].

В таблицах 1-3 приведены дисперсии оценок коэффициентов регрессионного полинома, рассчитанные по формулам, приведенным в главе 6 [1].

Табл. 1. Дисперсии оценок коэффициентов регрессионного полинома (10 входных точек)

Интервал	0	1	2	3	4	5
[10;20]	$8{,}4\times10^{6}\sigma^2$	$9{,}5\times10^{5}\sigma^2$	$1{,}7\times10^{4}\sigma^2$	$7{,}3\times10\sigma^2$	$7{,}7\times10^{-2}\sigma^2$	$1{,}3\times10^{-5}\sigma^2$
[10;1010]	$5\times10\sigma^2$	$1\times10^{-2}\sigma^2$	$2{,}4\times10^{-7}\sigma^2$	$1{,}1\times10^{-12}\sigma^2$	$1\times10^{-18}\sigma^2$	$1{,}3\times10^{-25}\sigma^2$
[0;10]	$3{,}7\times10\sigma^2$	$8{,}5\times10\sigma^2$	$2{,}1\times10\sigma^2$	$1\sigma^2$	$9{,}8\times10^{-3}\sigma^2$	$1{,}3\times10^{-5}\sigma^2$
[0;1000]	$3{,}7\times10\sigma^2$	$8{,}5\times10^{-3}\sigma^2$	$2{,}1\times10^{-7}\sigma^2$	$1\times10^{-12}\sigma^2$	$9{,}8\times10^{-19}\sigma^2$	$1{,}3\times10^{-25}\sigma^2$
[-5;5]	$4\times10^{-1}\sigma^2$	$2{,}5\times10^{-1}\sigma^2$	$4{,}3\times10^{-2}\sigma^2$	$7{,}5\times10^{-3}\sigma^2$	$1{,}4\times10^{-4}\sigma^2$	$1{,}3\times10^{-5}\sigma^2$
[-500;500]	$4\times10^{-1}\sigma^2$	$2{,}5\times10^{-5}\sigma^2$	$4{,}3\times10^{-10}\sigma^2$	$7{,}5\times10^{-15}\sigma^2$	$1{,}4\times10^{-20}\sigma^2$	$1{,}3\times10^{-25}\sigma^2$

Табл. 2. Дисперсии оценок коэффициентов регрессионного полинома (100 входных точек)

Интервал	0	1	2	3	4	5
[10;20]	$3{,}3\times10^{5}\sigma^2$	$4\times10^{4}\sigma^2$	$7{,}7\times10^{2}\sigma^2$	$3{,}5\sigma^2$	$4\times10^{-3}\sigma^2$	$7\times10^{-7}\sigma^2$
[10;1010]	$6{,}1\times10^{-1}\sigma^2$	$2{,}1\times10^{-4}\sigma^2$	$7{,}1\times10^{-9}\sigma^2$	$4{,}2\times10^{-14}\sigma^2$	$4{,}7\times10^{-20}\sigma^2$	$7\times10^{-27}\sigma^2$
[0;10]	$4{,}3\times10^{-1}\sigma^2$	$1{,}7\sigma^2$	$6{,}1\times10^{-1}\sigma^2$	$3{,}8\times10^{-2}\sigma^2$	$4{,}5\times10^{-4}\sigma^2$	$7\times10^{-7}\sigma^2$
[0;1000]	$4{,}3\times10^{-1}\sigma^2$	$1{,}7\times10^{-4}\sigma^2$	$6{,}1\times10^{-9}\sigma^2$	$3{,}8\times10^{-14}\sigma^2$	$4{,}5\times10^{-20}\sigma^2$	$7\times10^{-27}\sigma^2$
[-5;5]	$3{,}5\times10^{-2}\sigma^2$	$2{,}3\times10^{-2}\sigma^2$	$2{,}2\times10^{-3}\sigma^2$	$5{,}7\times10^{-4}\sigma^2$	$4{,}5\times10^{-6}\sigma^2$	$7\times10^{-7}\sigma^2$
[-500;500]	$3{,}5\times10^{-2}\sigma^2$	$2{,}3\times10^{-6}\sigma^2$	$2{,}2\times10^{-11}\sigma^2$	$5{,}7\times10^{-16}\sigma^2$	$4{,}5\times10^{-22}\sigma^2$	$7\times10^{-27}\sigma^2$

Табл. 3. Дисперсии оценок коэффициентов регрессионного полинома (1000 входных точек)

Интервал	0	1	2	3	4	5
[10;20]	$3{,}2\times10^{4}\sigma^2$	$3{,}9\times10^{4}\sigma^2$	$7{,}5\times10\sigma^2$	$3{,}5\times10^{-1}\sigma^2$	$3{,}9\times10^{-4}\sigma^2$	$7\times10^{-8}\sigma^2$
[10;1010]	$5{,}2\times10^{-2}\sigma^2$	$1{,}9\times10^{-5}\sigma^2$	$6{,}6\times10^{-10}\sigma^2$	$4\times10^{-15}\sigma^2$	$4{,}6\times10^{-21}\sigma^2$	$7\times10^{-28}\sigma^2$
[0;10]	$3{,}7\times10^{-2}\sigma^2$	$1{,}5\times10^{-1}\sigma^2$	$5{,}7\times10^{-2}\sigma^2$	$3{,}6\times10^{-3}\sigma^2$	$4{,}4\times10^{-5}\sigma^2$	$7\times10^{-8}\sigma^2$
[0;1000]	$3{,}7\times10^{-2}\sigma^2$	$1{,}5\times10^{-5}\sigma^2$	$5{,}7\times10^{-10}\sigma^2$	$3{,}6\times10^{-15}\sigma^2$	$4{,}4\times10^{-21}\sigma^2$	$7\times10^{-28}\sigma^2$
[-5;5]	$3{,}5\times10^{-3}\sigma^2$	$2{,}3\times10^{-3}\sigma^2$	$2{,}2\times10^{-4}\sigma^2$	$5{,}7\times10^{-5}\sigma^2$	$4{,}4\times10^{-7}\sigma^2$	$7\times10^{-8}\sigma^2$
[-500;500]	$3{,}5\times10^{-3}\sigma^2$	$2{,}3\times10^{-7}\sigma^2$	$2{,}2\times10^{-12}\sigma^2$	$5{,}7\times10^{-17}\sigma^2$	$4{,}4\times10^{-23}\sigma^2$	$7\times10^{-28}\sigma^2$

Как видно из таблиц 1-3, чем шире интервал, тем меньше дисперсии оценок коэффициентов. Также, дисперсии уменьшаются, когда интервал эксперимента симметричный точке ноль. Если невозможно принимать отрицательные значения, то рекомендуется начинать интервал в точке ноль. Интервал довольно сильно влияет на дисперсии оценок коэффициентов, даже больше чем изменение количества входных точек. Дисперсии оценок коэффициентов могут быть пересчитаны при масштабировании по формуле (32):

$$D\hat{\theta}_j^z = \sigma^2 \sum_{l=r}^{j} \left(\frac{1}{k^j} q_{lj}^x\right)^2 = \left(\frac{1}{k^j}\right)^2 D\hat{\theta}_j^x$$

С учетом того, что на практике количество входных точек и необходимая степень регрессионного полинома могут быть достаточно большими числами, это приводит к выполнению значительного количества расчетов. Перерасчет коэффициентов является достаточно простым процессом и не требует значительных компьютерных мощностей. Рассмотрим перерасчет дисперсий на примере (табл. 4).

Табл. 4. Пример перерасчета дисперсий оценок коэффициентов (1000 входных точек)

	[-5;5]	[-500;500] $k=100$
0	$3{,}5\times10^{-3}\sigma^2$	$\left(\frac{1}{100^0}\right)^2 \times 3{,}5\times10^{-3}\sigma^2 = 3{,}5\times10^{-3}\sigma^2$
1	$2{,}3\times10^{-3}\sigma^2$	$\left(\frac{1}{100^1}\right)^2 \times 2{,}3\times10^{-3}\sigma^2 = 2{,}3\times10^{-7}\sigma^2$
2	$2{,}2\times10^{-4}\sigma^2$	$\left(\frac{1}{100^2}\right)^2 \times 2{,}2\times10^{-4}\sigma^2 = 2{,}2\times10^{-12}\sigma^2$
3	$5{,}7\times10^{-5}\sigma^2$	$\left(\frac{1}{100^3}\right)^2 \times 5{,}7\times10^{-5}\sigma^2 = 5{,}7\times10^{-17}\sigma^2$
4	$4{,}4\times10^{-7}\sigma^2$	$\left(\frac{1}{100^4}\right)^2 \times 4{,}4\times10^{-7}\sigma^2 = 4{,}4\times10^{-23}\sigma^2$
5	$7\times10^{-8}\sigma^2$	$\left(\frac{1}{100^5}\right)^2 \times 7\times10^{-8}\sigma^2 = 7\times10^{-28}\sigma^2$

Таким образом, можно заранее увидеть эффективность регрессионного анализа, и подобрать необходимые исследователю диапазоны. Аналогично, по формуле (16) можно пересчитать значения коэффициентов нормированных ортогональных полиномов Форсайта: что требует значительно меньше вычислений, чем расчет этих значений заново по формулам (3)-(8).

$$q^z_{jl} = \frac{1}{k^l} q^x_{jl}, \forall j = \overline{0,r}, \forall l = \overline{0,j}$$

Такой пересчет требует значительно меньше вычислений, чем расчет этих значений заново по формулам (3)-(8).

Однако, проведение масштабирования уже полученных входных данных не приведет к улучшению результата, дальше это будет доказано.

3. Масштабирование входных данных

Как показали эксперименты, при масштабировании уже имеющихся входных данных, качество регрессионного полинома не изменится. То есть, несмотря на то, что изменятся оценки коэффициентов и их дисперсии, значение регрессионного полинома при заданном аргументе останется таким же. Для доказательства этого используем формулы из главы 6 [1]. Оценки коэффициентов регрессионного полинома получаются с помощью нормированных ортогональных полиномов Форсайта.

$$\hat{\theta}_j = \hat{w}_r q_{rj} + \cdots + \hat{w}_j q_{jj}, j = \overline{0,r} \qquad (1)$$

В формуле (1) $\hat{\theta}_j$ – оценки коэффициентов регрессионного полинома, $\hat{w}_j$ – оценки весовых коэффициентов, полученные методом наименьших квадратов, q_{jj} – коэффициенты нормированных ортогональных полиномов Форсайта. Расчет оценок весовых коэффициентов происходит по формуле (2).

$$\hat{w}_j = \sum_{i=1}^{n} y_i \theta_j(x_i), j = \overline{0,r}, \qquad (2)$$

где $\theta_j(x_i)$ – значение $j^{ого}$ нормированного ортогонального полинома Форсайта на аргументе x_i.

Коэффициенты нормированных ортогональных полиномов Форсайта рассчитываются по рекуррентной формуле (3) из двух предыдущих.

$$\lambda\theta_j(x) = x\theta_{j-1}(x) - \alpha\theta_{j-1}(x) - \beta\theta_{j-2}(x), \qquad (3)$$

где α, β, λ – коэффициенты, которые рассчитываются по формулам (4), (5) и (6) соответственно.

$$\alpha = \sum_{i=1}^{n} x_i \theta^2_{j-1}(x_i) \qquad (4)$$

$$\beta = \sum_{i=1}^{n} x_i \theta_{j-1}(x_i)\theta_{j-2}(x_i) \qquad (5)$$

$$\lambda = \sqrt{\sum_{i=1}^{n} (x_i\theta_{j-1}(x_i) - \alpha\theta_{j-1}(x_i) - \beta\theta_{j-2}(x_i))^2} \qquad (6)$$

Первых два ортогональных полинома рассчитываются по формулам (7) и (8) соответственно.

$$\theta_0(x) = \frac{1}{\sqrt{n}} \qquad (7)$$

$$\theta_1(x) = -\frac{\bar{x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} + \frac{x}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (8)$$

Для доказательства того, что при масштабировании качество входных данных регрессионного полинома не изменится, будем рассматривать коэффициенты нормированных ортогональных полиномов Форсайта отдельно. Тогда из (3), (7) и (8) получаем выражения (9) - (15).

$$q_{00} = \frac{1}{\sqrt{n}} \qquad (9)$$

$$q_{10} = -\frac{\bar{x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (10)$$

$$q_{11} = \frac{1}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (11)$$

$$q_{j0} = \frac{\alpha q_{j-1,0} - \beta q_{j-2,0}}{\lambda} \qquad (12)$$

$$q_{jl} = \frac{q_{j-1,l-1} - \alpha q_{j-1,l} - \beta q_{j-2,l}}{\lambda}, l = \overline{1, j-2} \qquad (13)$$

$$q_{j,j-1} = \frac{q_{j-1,j-2} - \alpha q_{j-1,j-1}}{\lambda} \qquad (14)$$

$$q_{j,j} = \frac{q_{j-1,j-1}}{\lambda} \qquad (15)$$

Формулы (12) - (15) выполняются $\forall j = \overline{2,r}$.

Предположим, что было выполнено масштабирование исходных данных следующим образом $z = kx$, где k – некоторый коэффициент масштабирования, который является действительным числом. Обозначим коэффициенты при начальном масштабировании с помощью

q^x, а при новом – q^z. Покажем, что связь коэффициентов нормированных ортогональных полиномов при разном масштабировании соответствует формуле (16).

$$q^z_{jl} = \frac{1}{k^l} q^x_{jl}, \forall j = \overline{0,r}, \forall l = \overline{0,j} \qquad (16)$$

Рассмотрим связь между средним значением аргументов начального и нового масштабирования, поскольку оно используется в формулах (10) и (11).

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} kx_i = k\bar{x} \qquad (17)$$

Рассмотрим коэффициенты первых двух нормированных ортогональных полиномов Форсайта

$$q^z_{00} = \frac{1}{\sqrt{n}} = q^x_{00} \qquad (18)$$

$$q^z_{10} = \frac{k\bar{x}}{\sqrt{\sum_{i=1}^{n}(kx_i - k\bar{x})^2}} = q^x_{10} \qquad (19)$$

$$q^z_{11} = \frac{1}{\sqrt{\sum_{i=1}^{n}(kx_i - k\bar{x})^2}} = \frac{1}{k} q^x_{11} \qquad (20)$$

Следовательно, для коэффициентов первых двух ортогональных полиномов выполняется формула (16). Рассмотрим значения ортогонального полинома при заданном аргументе, для которых выполняется формула (16).

$$\theta^z_j(z_i) = q^x_{j0} + \cdots + \frac{1}{k^j} q^x_{jj} k^j x_i = \theta^x_j(x_i) \qquad (21)$$

Рассмотрим связь коэффициентов α, β, λ при различных масштабированиях.

$$\alpha^z = \sum_{i=1}^{n} kx_i \theta^2_{j-1}(x_i) = k\alpha^x \qquad (22)$$

$$\beta^z = \sum_{i=1}^{n} kx_i \theta_{j-1}(x_i)\theta_{j-2}(x_i) = k\beta^x \qquad (23)$$

$$\lambda^z = \sqrt{\sum_{i=1}^{n}(kx_i\theta^x_{j-1}(x_i) - k\alpha\theta^x_{j-1}(x_i) - k\beta\theta^x_{j-2}(x_i))^2} = k\lambda^x \qquad (24)$$

Рассмотрим связь между коэффициентами полиномов $j = \overline{2,r}$ при различных масштабированиях.

$$q^z_{j0} = \frac{k\alpha^x q^x_{j-1,0} - k\beta^x q^x_{j-2,0}}{k\lambda^x} = q^x_{j0} \qquad (25)$$

$$q^z_{jl} = \frac{\frac{1}{k^{l-1}} q^x_{j-1,l-1} - k\alpha^x \frac{1}{k^l} q^x_{j-1,l} - k\beta^x \frac{1}{k^l} q^x_{j-2,l}}{k\lambda^x} = \frac{1}{k^l} q^x_{jl}, l = \overline{1, j-2} \qquad (26)$$

$$q^z_{j,j-1} = \frac{\frac{1}{k^{j-2}} q^x_{j-1,j-2} - k\alpha^x \frac{1}{k^{j-1}} q^x_{j-1,j-1}}{k\lambda^x} = \frac{1}{k^{j-1}} q^x_{j,j-1} \qquad (27)$$

$$q^z_{j,j} = \frac{\frac{1}{k^{j-1}} q^x_{j-1,j-1}}{k\lambda^x} = \frac{1}{k^j} q^x_{j,j} \qquad (28)$$

Итак, для ортогональных полиномов Форсайта формула (16) справедлива, при масштабировании $z = kx$.

Рассмотрим связь оценок весовых коэффициентов при различных масштабах.

$$\hat{w}^z_j = \sum_{i=1}^{n} y_i \theta^x_j(x_i) = \hat{w}^x_j \qquad (29)$$

Исходя из формул (16) и (29), связь между оценками коэффициентов регрессионного полинома при различных масштабированиях будет следующая.

$$\hat{\theta}^z_j = \hat{w}_r \frac{1}{k^j} q^x_{rj} + \cdots + \hat{w}_j \frac{1}{k^j} q^x_{jj} = \frac{1}{k^j}\hat{\theta}^x_j \qquad (30)$$

Исходя из (30), связь между значениями регрессионного полинома при заданном аргументе и различных масштабированиях будет следующая.

$$f^z(z_i) = \sum_{j=0}^{r} \frac{1}{k^j} \hat{\theta}^x_j k^j x_i = f^x(x_i) \qquad (31)$$

Следовательно, значение регрессионного полинома, при различных масштабированиях имеющихся входных данных останется одинаковым, хотя вид самого регрессионного полинома изменяется согласно формуле (30).

Рассмотрим связь между дисперсиями оценок коэффициентов регрессионного полинома при различных масштабированиях.

$$D\hat{\theta}_j^z = \sigma^2 \sum_{l=r}^{j} \left(\frac{1}{k^j} q_{lj}^x\right)^2 = \left(\frac{1}{k^j}\right)^2 D\hat{\theta}_j^x \qquad (32)$$

Следовательно, дисперсии оценок коэффициентов регрессионного полинома изменяются по формуле (32).

4. Вывод

При проведении активного эксперимента с получением входных данных регрессии, дисперсии оценок коэффициентов регрессионного полинома будут зависеть от интервала, на котором он проводится. Чем ближе центр интервала до точки ноль, тем меньшими будут дисперсии. Однако значительного снижения дисперсий можно достичь путем расширения интервала. Однако, при масштабировании уже имеющихся входных данных, качество регрессионного полинома не изменится, хотя изменится его коэффициенты и их дисперсии.

Список литературы

1. М.З. Згуровский, А.А. Павлов Принятие решений в сетевых системах с ограниченными ресурсами. – К.: Наукова думка, – 2010. – 569 с.