

Тема X2. Обработка и анализ данных с интервальной неопределённостью.

А.Н. Баженов

ФТИ им. А.Ф.Иоффе

a_bazhenov@inbox.ru

26.04.2021

Обработка и анализ данных с интервальной неопределённостью.

ПЛАН

- Общие понятия
- Обработка константы
- Задача восстановления зависимостей

Теория:

А.Н. БАЖЕНОВ, С.И. ЖИЛИН, С.И. КУМКОВ, С.П. ШАРЫЙ.
Обработка и анализ данных с интервальной неопределённостью. РХД.
Серия «Интервальный анализ и его приложения». Ижевск. 2021. с.200.

Задача восстановления зависимостей. Часть 1.

Задача восстановления зависимостей

Даются определения новых терминов и понятий, которые возникают в связи с восстановлением функциональных зависимостей по данным их измерений и наблюдений, имеющих интервальную неопределённость.

Мы рассмотрим основные идеи и типичные приёмы восстановления зависимостей по интервальным данным, а также возникающие при этом проблемы.

Подробно исследуется случай простейшей линейной зависимости, но большинство построений и рассуждений легко переносятся на общий нелинейный случай.

Постановка задачи

Предположим, что величина y является функцией некоторого заданного вида от независимых аргументов x_1, x_2, \dots, x_m , т. е.

$$y = f(x, \beta), \quad (1)$$

где $x = (x_1, \dots, x_m)$ — вектор независимых переменных, $\beta = (\beta_1, \dots, \beta_l)$ — вектор параметров функции. Имея набор значений переменных x и y , нам нужно найти β_1, \dots, β_l , которые соответствуют конкретной функции f из параметрического семейства (1).

Мы будем называть эту задачу *задачей восстановления зависимости*.

Постановка задачи

Важнейший частный случай поставленной задачи — определение параметров линейной функциональной зависимости вида

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m, \quad (2)$$

в которой x_1, x_2, \dots, x_m — независимые переменные (которые называются также *экзогенными*, *предикторными* или просто *входными* переменными), y — это зависимая переменная (которая называется также *эндогенной*, *критериальной* или *выходной* переменной), а $\beta_0, \beta_1, \dots, \beta_m$ — некоторые коэффициенты.

Эти неизвестные коэффициенты должны быть определены из ряда измерений значений x_1, x_2, \dots, x_m и y .

Постановка задачи

Результаты измерений неточны, и мы предполагаем что они имеют *ограниченную неопределённость*, когда нам известны лишь некоторые интервалы, дающие двусторонние границы измеренных значений.

Таким образом, результатом i -го измерения являются такие интервалы $\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \dots, \mathbf{x}_m^{(i)}, \mathbf{y}^{(i)}$, относительно которых мы предполагаем, что истинное значение x_1 лежит в пределах $\mathbf{x}_1^{(i)}$, истинное значение x_2 лежит в $\mathbf{x}_2^{(i)}$ и т.д. вплоть до y , истинное значение которого находится в интервале $\mathbf{y}^{(i)}$.

В целом имеется n измерений, так что индекс i может принимать значения из множества натуральных чисел $\{1, 2, \dots, n\}$.

Постановка задачи

Далее для удобства построений и выкладок обозначим номер измерения i не верхним, а нижним индексом, который мы поставим первым при обозначении входов. Таким образом, полный набор данных будет иметь вид

$$\begin{array}{ccccc} x_{11}, & x_{12}, & \dots & x_{1m}, & y_1, \\ x_{21}, & x_{22}, & \dots & x_{2m}, & y_2, \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1}, & x_{n2}, & \dots & x_{nm}, & y_n. \end{array} \quad (3)$$

Нам необходимо найти или как-то оценить коэффициенты β_j , $j = 0, 1, \dots, m$, для которых линейная функция (2) «наилучшим образом» приближала бы интервальные данные измерений (3).

Для обозначения $n \times m$ -матрицы, составленной из данных (3) для независимых переменных часто используют термины *матрица плана эксперимента* или просто *матрица плана*, которые возникли в теории планирования эксперимента .

Интервалы $x_{i1}, x_{i2}, \dots, x_{im}, y_i$ мы называем, как и раньше, *интервалами неопределённости i -го измерения*.

Но кроме них нам также потребуется обращаться ко всему множеству, ограничиваемому в многомерном пространстве \mathbb{R}^{m+1} этими интервалами по отдельным координатным осям.

Definition

Брусом неопределённости i -го измерения рассматриваемой зависимости будем называть интервальный вектор-брус $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{im}, \mathbf{y}_i) \subset \mathbb{R}^{m+1}$, $i = 1, 2, \dots, n$.

Таким образом, каждый брус неопределённости измерения зависимости является прямым декартовым произведением интервалов неопределённости независимых переменных и зависимой переменной. На Рис. 1 на плоскости Oxy наглядно показаны брусы неопределённости измерений и график линейной функции, которую мы восстанавливаем.

Далее мы рассматриваем данные (3) как «спущенные свыше» и никак не обсуждаем их выбор, коррекцию или оптимизацию.

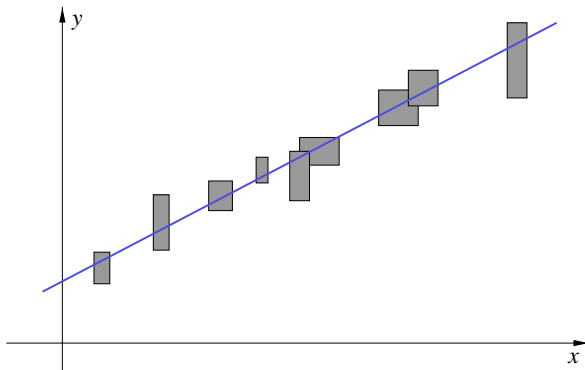


Рис.: Наглядная иллюстрация задачи восстановления линейной зависимости по данным с интервальной неопределённостью.

Definition

Будем называть брус неопределённости измерения зависимости *накрывающим*, если он гарантированно содержит истинные значения измеряемых величин входных и выходных переменных зависимости.

Брус неопределённости измерения зависимости, который не является накрывающим, будем называть *ненакрывающим*.

Возможные альтернативные термины — «включающий брус неопределённости», «охватывающий брус неопределённости» (их отрицание — «невключающий», «неохватывающий»).

Диаграммы рассеяния

Для визуализации интервальных данных, аналогично традиционному точечному случаю, используют *диаграммы рассеяния*.

В традиционном понимании диаграмма рассеяния используется в статистике и анализе данных для визуализации значений двух переменных в виде «облака» точек на декартовой плоскости и позволяет оценить наличие или отсутствие корреляции и других взаимосвязей между двумя переменными.

На диаграмме рассеяния для интервальных данных каждое интервальное наблюдение отображается в виде бруса (бруса неопределённости). При отсутствии неопределённости по одной из переменных, брусы наблюдений могут «схлопываться» в одномерные вертикальные или горизонтальные отрезки («ворота»).

Примерами диаграмм рассеяния могут служить Рис. 1 и Рис. 3.

Definition

Накрывающая выборка — совокупность накрывающих измерений, т. е. выборка, в которой все измерения (наблюдения) являются накрывающими.

Напротив, выборка называется *ненакрывающей*, если хотя бы одно из входящих в неё измерений — ненакрывающее.

Решение задачи восстановления зависимостей для обычных точечных данных

Существует большое количество более или менее стандартных подходов к решению задачи восстановления зависимостей для обычных точечных данных.

Наиболее популярные из них — это метод наименьших квадратов, метод наименьших модулей и метод максимальной энтропии. Часто используется чебышёвское (минимаксное) сглаживание.

Все эти методы основаны на нахождении глобального (абсолютного) минимума определённым образом подобранной целевой функции. Мы пытаемся найти наиболее набор параметров, который доставляет минимум этому функционалу. Очевидно, что конечный результат будет существенно отличаться в зависимости от формы этого целевого функционала.

В любом случае, «идеальным решением» задачи можно считать ту функциональная зависимость вида (если она существует), линия графика которой проходит через все точки данных.

Что следует считать решением?

Что следует считать решением задачи восстановления зависимости по интервальному данным (3)?

Очевидно, что функцию, вида (1) или (2), нужно считать точным решением задачи восстановления искомой зависимости, если её график проходит через все брусы неопределённости данных.

В случае точечных данных эта идеальная ситуация почти никогда не реализуется и неустойчива к малым возмущениям в данных. Но в случае данных с существенной интервальной неопределённостью прохождение графика функции через брусы данных (3) может реализовываться, и оно устойчиво к возмущениям в данных.

Кроме того, дополнительную специфику задаче придаёт то новое обстоятельство, что брусы неопределённости данных (3), в отличие от бесконечно малых и бесструктурных точек, получают структуру и потому нужно различать, как именно проходит график функции через эти брусы.

В соответствии с терминологией, намеченной для нахождения констант, будем называть *информационным множеством* задачи восстановления зависимости множество значений параметров зависимости, совместных с данными в каком-то определённом смысле.

В традиционном «точечном» случае, когда данные неинтервальны, решение задачи восстановления зависимостей получается по следующей общей схеме. Мы подставляем данные в формулу для зависимости (2) и получаем для каждого отдельного измерения одно уравнение. В целом в результате этой процедуры возникает система уравнений, решив которую, в обычном или обобщённом смысле, мы найдём параметры зависимости.

В интервальном случае, действуя аналогичным образом, мы получим уже интервальную систему уравнений, которую также можно решать. Её решением, обычным или в некотором обобщённом смысле, будет вектор оценки параметров восстанавливаемой зависимости (2).

Информационное множество задачи получается при этом как множество решений этой интервальной системы уравнений, построенной на основе формулы (2) и данных (3).

Определение параметров функциональной зависимости производится, как правило, для того, чтобы затем найденную формулу использовать для предсказания значений зависимости в других интересующих нас точках её области определения.

Ясно, что такое предсказание будет осуществляться с некоторой погрешностью, вызванной неопределённостями данных, неоднозначностью самой процедуры восстановления и т. п. Эту неопределённость предсказания также необходимо знать и учитывать в нашей деятельности.

Коридор совместных зависимостей и его сечение

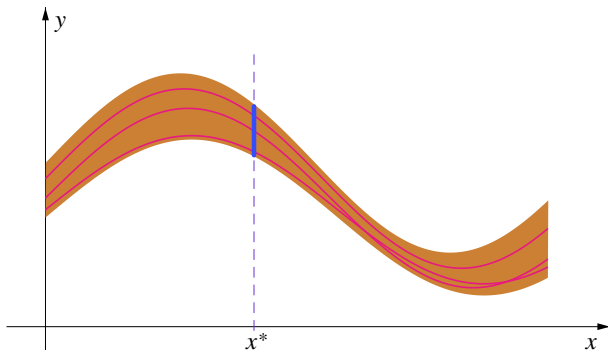


Рис.: Коридор совместных зависимостей и его сечение
для какого-то значения аргумента x^* .

Коридор совместных зависимостей

Если информационное множество задачи восстановления зависимостей непусто, то обычно оно задаёт целое семейство зависимостей, совместных с данными задачи, которое имеет смысл рассматривать вместе, как единое целое.

Это необходимо делать в вопросах, касающихся оценивания неопределённости предсказания, учёта всех возможных сценариев развития и т. п. Как следствие, возникает необходимость рассматривать вместе, единым целым, множество всех функций, совместных с интервальными данными задачи восстановления зависимости. Мы будем называть его *коридором совместных зависимостей* (см. Рис. 2).

Многозначные отображения

В литературе использовались также другие термины для обозначения этого объекта — «трубка» совместных зависимостей (имеет происхождение в теории управления), «полоса» или даже «слой неопределённости», «коридор неопределённости» и т. п.

Строгое определение коридора совместных зависимостей может быть дано на основе математического понятия многозначного отображения. Напомним, что для произвольных множеств X и Y *многозначным отображением* F из X в Y называется соответствие (правило), сопоставляющее каждой точке $x \in X$ непустое подмножество $F(x) \subset Y$, называемое *значением* или *образом* x .

Definition

Пусть в задаче восстановления зависимостей информационное множество Ω параметров зависимостей $y = f(x, \beta)$, совместных с данными, является непустым. *Коридором совместных зависимостей* рассматриваемой задачи называется многозначное отображение \mathcal{Y} , сопоставляющее каждому значению аргумента x множество

$$\mathcal{Y}(x) = \bigcup_{\beta \in \Omega} f(x, \beta).$$

Сечение коридора совместных зависимостей

Значение $\mathcal{Y}(\tilde{x})$ коридора совместных зависимостей при каком-то определённом аргументе \tilde{x} («сечение коридора») — это множество $\bigcup_{\beta \in \Omega} f(\tilde{x}, \beta)$, образованное всевозможными значениями, которые принимают на этом аргументе функциональные зависимости, совместные с интервальными данными измерений.

Рис. 2 изображает коридор совместных зависимостей в задаче восстановления нелинейной зависимости, но для рассматриваемого нами линейного случая коридор совместных значений имеет существенно более специальный вид .

Нетрудно показать, что границы коридора совместных зависимостей в этом случае являются *кусочно-линейными*.

Случай точных измерений входных переменных

Важнейшим и часто встречающимся частным случаем рассмотренной задачи является ситуация, когда независимые (экзогенные, предикторные, входные) переменные x_1, x_2, \dots, x_m измеряются точно, и вместо телесных брусков неопределённости измерений (как на Рис. 1) мы имеем отрезки прямых $(x_{i1}, x_{i2}, \dots, x_{im}, y_i)$, $i = 1, 2, \dots, n$, параллельные оси зависимой (эндогенной, критериальной, выходной) переменной (см. Рис. 3).

Именно такая постановка задачи была рассмотрена в пионерской работе Л.В. Канторовича.

Случай точных измерений входных переменных

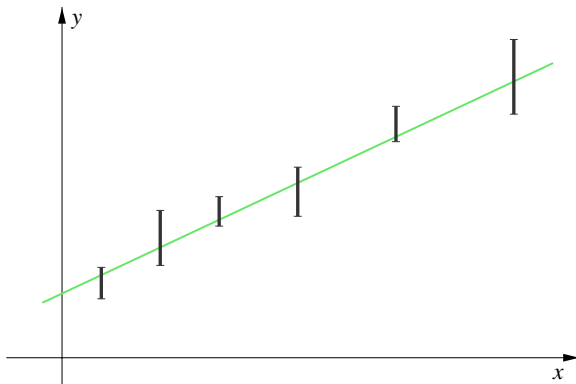


Рис.: Частный случай задачи восстановления линейной зависимости по неточным данным, когда входные переменные измеряются точно.

Отсутствие неопределённости значений независимых переменных приводит к кардинальному упрощению математической модели. Брусы неопределённости измерений зависимости, введённые ранее, схлопываясь по независимым переменным, превращаются в *отрезки неопределённости*.

Как следствие, для решения и полного исследования этого частного случая предложено большое количество эффективных вычислительных методов. Рассмотрим эти математические вопросы более детально.

Линейная зависимость (2) *совместна* (согласуется) с интервальными данными измерений, если её график проходит через все отрезки неопределённости, задаваемые интервалами измерений выходной переменной y , как это изображено на Рис. 3).

Подобное понимание совместности (согласования) является прямым обобщением того понимания «совместности», которое традиционно для неинтервального случая и используется, к примеру в постановке задачи интерполяции.

Подставляя в зависимость (2) данные для входных переменных x_1, x_2, \dots, x_m в i -ом измерении и требуя включения полученного значения в интервалы y_i , получим

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in y_i, \quad i = 1, 2, \dots, n. \quad (4)$$

Фактически, это интервальная система линейных алгебраических уравнений

$$\begin{cases} \beta_0 + x_{11}\beta_1 + x_{12}\beta_2 + \dots + x_{1m}\beta_m = y_1, \\ \beta_0 + x_{21}\beta_1 + x_{22}\beta_2 + \dots + x_{2m}\beta_m = y_2, \\ \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \ddots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ \beta_0 + x_{n1}\beta_1 + x_{n2}\beta_2 + \dots + x_{nm}\beta_m = y_n, \end{cases}$$

у которой интервальность присутствует только в правой части.

С другой стороны, (4) равносильно системе

$$\left\{ \begin{array}{l} \underline{y}_1 \leq \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_m x_{1m} \leq \overline{y}_1, \\ \underline{y}_2 \leq \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_m x_{2m} \leq \overline{y}_2, \\ \vdots \quad \vdots \quad \quad \quad \ddots \quad \quad \quad \vdots \quad \vdots \\ \underline{y}_n \leq \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_m x_{nm} \leq \overline{y}_n. \end{array} \right. \quad (5)$$

Система двусторонних линейных неравенств

Это система двусторонних линейных неравенств относительно неизвестных параметров $\beta_0, \beta_1, \beta_2, \dots, \beta_m$, решив которую, мы можем найти искомую линейную зависимость. Множество решений системы неравенств (5) естественно считать информационным множеством параметров восстанавливаемой зависимости для рассматриваемого случая.

Для i -го двустороннего неравенства из системы (5) множество решений — это полоса в пространстве \mathbb{R}^{m+1} параметров $(\beta_0, \beta_1, \dots, \beta_m)$, ограниченная с двух сторон гиперплоскостями с уравнениями

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} = \underline{y}_i,$$

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} = \overline{y}_i.$$

Система двусторонних линейных неравенств

Множество решений системы неравенств (5) является пересечением n штук таких полос, отвечающих отдельным измерениям. Можно рассматривать эти полосы как информационные множества отдельных измерений.

На Рис. 4 изображено формирование множества решений системы неравенств (5) для случая двух параметров (т. е. $m = 1$) и $n = 3$.

Образование информационного множества параметров

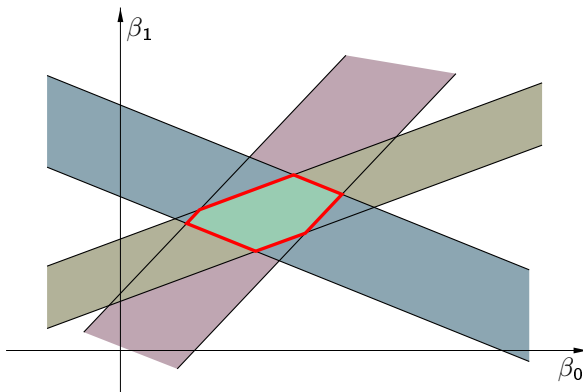


Рис.: Образование информационного множества параметров \mathcal{Y} линейной зависимости (ограничено красной линией) для случая точных входных переменных.

Информационное множество — трудоёмкость распознавания

В целом множество решений системы линейных алгебраических неравенств (5) является *выпуклым многогранным множеством в пространстве \mathbb{R}^{m+1}* .

Распознавание того, пусто оно или непусто, а также нахождение какой-либо точки из него, являются задачами, сложность которых ограничена полиномом от их размера. Существуют эффективные и хорошо разработанные вычислительные методы для решения этих вопросов и для нахождения оценок множества решений, например, основанные на сведении рассматриваемой задачи к задаче линейного программирования.

Информационное множество — трудоёмкость распознавания

В общем случае, когда входные (экзогенные, предикторные) переменные известны неточно, ситуация существенно усложняется и множество параметров, совместных (согласующихся) с интервальными данными не может быть описано так же просто, с помощью системы линейных неравенств (5).

Трудоёмкость распознавания его пустоты или непустоты также становится экспоненциальной в зависимости от количества переменных [3].

Случай точных измерений входных переменных

Общий случай задачи восстановления зависимостей

Рассмотрим теперь случай, когда неопределённость присутствует как в измерениях значений зависимой переменной, так и в измерениях значений аргументов.

Это может быть вызвано различными причинами. Например, существенно неточное измерение входных переменных происходит в ситуациях, когда они должны устанавливаться в течение значительного времени.

Тогда их уместно выразить какими-то интервалами, а не точечными значениями.

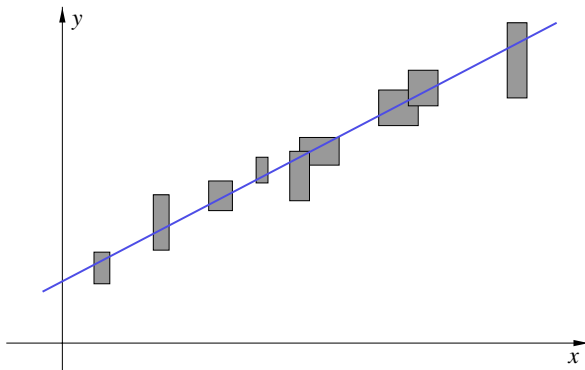


Рис.: Наглядная иллюстрация задачи восстановления линейной зависимости по данным с интервальной неопределённостью.

`https://github.com/szhilin/octave-interval-examples/blob/master/SteamGenerator.ipynb`

Общий случай задачи восстановления зависимостей

Это формальная запись, означающая совокупность обычных (точечных) систем линейных алгебраических уравнений того же размера и с теми же неизвестными переменными, у которых коэффициенты и правые части лежат в предписанных им интервалах (см. [3]).

Восстановление параметров линейной зависимости можно рассматривать как «решение», в том или ином смысле, выписанной интервальной системы уравнений.

Общий случай задачи восстановления зависимостей

В случае присутствия погрешностей как в измерениях аргумента, так и в измерениях зависимости множество параметров зависимостей, совместных (согласующихся) с данными, характеризуются новыми свойствами, которыми не обладают задачи с точными измерениями входных переменных.

Прежде всего, множества решений отдельных интервальных уравнений уже *не являются полосами в пространстве \mathbb{R}^n* , вроде тех, что изображены на Рис. 4. Они выглядят существенно иначе, и их конкретный вид зависит от того, какой смысл вкладывается в понятие совместности (согласования) параметров и данных, т. е. от того, *какое множество решений ИСЛАУ взято в качестве информационного множества* (см. Рис. 6).

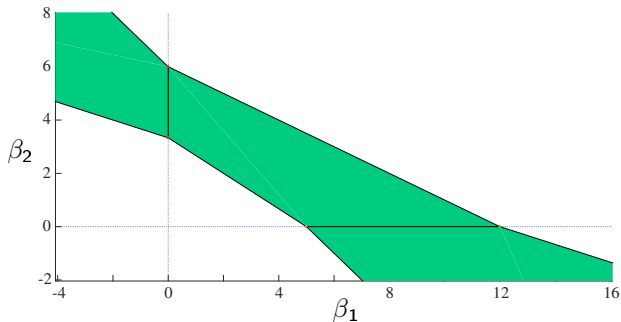


Рис.: Объединённое множество решений интервального
линейного уравнения $[1, 2]\beta_1 + [2, 3]\beta_2 = [10, 12]$.

Общий случай задачи восстановления зависимостей

Само понятие согласования (совместности) параметров и данных должно быть расширено и переосмыслено.

В обычном неинтервальном случае результаты измерений — это бесконечно малые точки, и прохождение через них графика функциональной зависимости адекватно описывается двумя значениями — «да» или «нет», т. е. имеет булевский (логический) тип данных.

Общий случай задачи восстановления зависимостей

Если мы переходим от точек к брусам неопределённости, то прохождение графика зависимости через них можно понимать по-разному.

Брусы неопределённости измерений являются прямыми декартовыми произведениями интервалов по различным осям координат, и эти оси имеют разный смысл:

интервалы $x_{i1}, x_{i2}, \dots, x_{im}$ соответствуют входным (экзогенным, предикторным) переменным, а интервал y_i соответствует выходной (эндогенной, критериальной) переменной.

По этой причине становится важным, как именно проходит график восстанавливаемой зависимости через брусы неопределённости измерений (см. Рис. 7).

Общий случай задачи восстановления зависимостей

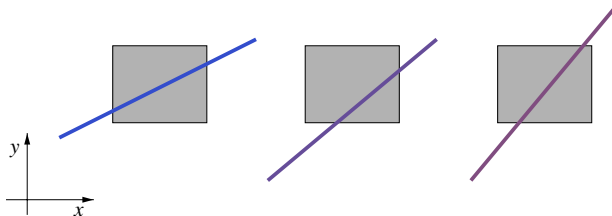


Рис.: Различные способы пересечения линии с бруском
неопределённости измерения зависимости.

Функциональную зависимость назовём *слабо совместной* с интервальными данными, если её график проходит через каждый брус неопределённости измерений хотя бы для одного значения аргумента.

Наглядно это означает, что график зависимости пересекает брусы неопределённости, но как именно — неважно (средний чертёж на Рис. 7),

достаточно лишь одной точки пересечения.

достаточно лишь одной точки пересечения.

Слабо совместная зависимость

Для случая линейной зависимости это условие наиболее удобно выразить с помощью формального языка логического исчисления предикатов:

$$(\exists x_{i1} \in \mathbf{x}_{i1}) \cdots (\exists x_{im} \in \mathbf{x}_{im})(\exists y_i \in \mathbf{y}_i) \\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in \mathbf{y}_i, \quad i = 1, 2, \dots, n.$$

Равносильная упрощённая формулировка этого свойства выглядит следующим образом:

$$(\exists x_{i1} \in \mathbf{x}_{i1}) \cdots (\exists x_{im} \in \mathbf{x}_{im}) \\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in \mathbf{y}_i, \quad i = 1, 2, \dots, n.$$

Сильно совместная зависимость

Функциональную зависимость назовём *сильно совместной* с интервальными данными, если её график проходит через каждый брус неопределённости измерений для любого значения аргумента из интервалов неопределённости входных переменных.

Наглядно это означает, что график зависимости

целиком содержится в коридорах,
задаваемых интервалами выходной переменной при всех значениях
входных переменных из соответствующих им интервалов

(левый чертёж на Рис. 7).

Для случая линейной зависимости это условие может быть формально записано в следующем виде:

$$(\forall x_{i1} \in \mathbf{x}_{i1}) \cdots (\forall x_{im} \in \mathbf{x}_{im})(\exists y_i \in \mathbf{y}_i) \\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in \mathbf{y}_i, \quad i = 1, 2, \dots, n.$$

Равносильная упрощённая формулировка этого свойства выглядит следующим образом:

$$(\forall x_{i1} \in \mathbf{x}_{i1}) \cdots (\forall x_{im} \in \mathbf{x}_{im}) \\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} \in \mathbf{y}_i, \quad i = 1, 2, \dots, n.$$

Сильно и слабо совместные зависимости

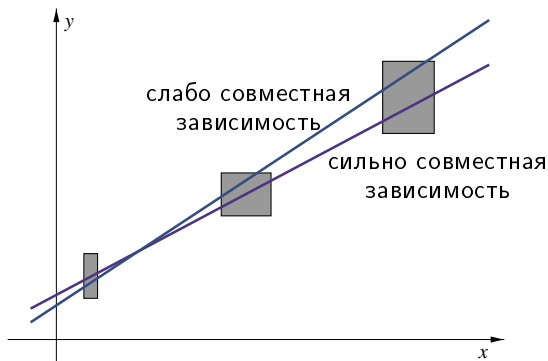


Рис.: Линейные зависимости с разными типами согласования с данными.

В чём содержательный смысл сильной совместности?

На практике измерения на входах и выходах системы осуществляются, как правило, разными способами и даже в разное время.

Мы измеряем выход (зависимую переменную) уже тогда, когда входные значения (независимых переменных) зафиксированы, и мы их измерили. Получив при этом какие-то интервалы.

Сильная совместность функциональной зависимости с интервальными данными означает тогда, что выходная величина остаётся в пределах измеренного для неё интервала вне зависимости от того, какими конкретно в своих интервалах являются значения входных переменных.

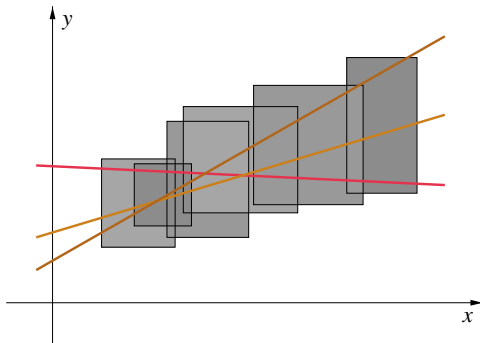


Рис.: Сложный случай восстановления зависимости по широким перекрывающимся интервальным данным.

Если матрица системы (6) уравнений — точечная, т. е. коэффициенты при неизвестных β_i являются обычными вещественными числами, то объединённое множество решений в целом является выпуклым.

Но в общем случае, когда матрица интервальной системы линейных алгебраических уравнений существенно интервальна, то объединённое множество решений может быть невыпуклым.

Допусковое множество решений всегда выпукло. В целом, количество гиперплоскостей, ограничивающих множества решений, может быть очень большим.

Возвращаясь к решению задачи восстановления зависимостей, следует отметить, что непростое строение множеств решений интервальных систем уравнений делает очень трудоёмким и малополезным их точное и полное описание.

Имеет смысл найти какое-нибудь приближённое описание информационного множества.

Здесь могут встретиться различные ситуации.

Приближённое описание информационного множества

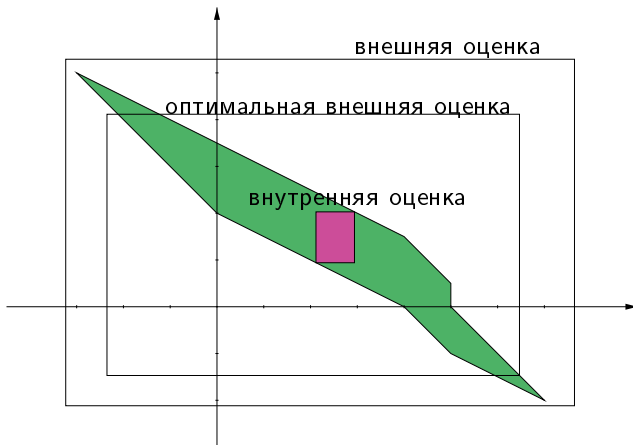


Рис.: Различные способы оценивания
информационного множества.

Оценки информационного множества

Часто бывает необходимо оценить разброс точек из информационного множества, то есть определить, насколько сильно оно «растекается» в пространстве параметров.

Часто это делается для его отдельных компонент, так что в целом нам требуется интервальный брус, содержащий множество решений. Это *внешняя оценка* информационного множества

Среди всех внешних оценок наилучшей служит минимальная по размерам внешняя оценка, которую также называют *оптимальной внешней оценкой*. Она единственна и является интервальной оболочкой информационного множества задачи.

Внешняя оценка информационного множества необходима, к примеру, при построении внешней оценки коридора совместных зависимостей, когда мы хотим просчитать гарантированный эффект от реализации всех сценариев, могущих встретиться по восстановленным зависимостям.

Оценки информационного множества

Во многих задачах требуется оценивание информационного множества с помощью какого-то несложно описываемого подмножества — *внутреннее оценивание*. Такая оценка будет содержать только точки из информационного множества и ничего лишнего.

Внешняя оценка информационного множества в этом смысле плоха тем, что включает в себя точки, не принадлежащие информационному множеству.

Если в качестве подмножества информационного множества берётся вписанный брус, то он называется *внутренней интервальной оценкой* множества решений. Среди двух внутренних оценок лучшей является та, которая целиком содержит другую, но максимальных по включению внутренних оценок, которые несравнимы друг с другом, может быть много.

Английские термины для обозначения внешней и внутренней оценки — outer estimate и inner estimate соответственно. Внешнюю оценку часто называют также термином «enclosure».

Кроме внешнего и внутреннего оценивания информационных множеств могут встретиться и другие, которые требуются по смыслу задачи.

Например, «слабое внешнее» оценивание, оценивание вдоль какого-то специального выделенного направления, исчерпывающее оценивание с помощью набора брусков и т.п.

Варианты точечной оценки информационного множества

Помимо оценивания информационного множества «целиком», во многих ситуациях достаточно найти какую-либо точку из него (здесь мы имеем аналогию с оцениванием «точечным» и «интервальным» в традиционной статистике). Естественно выбрать такую одну точку удовлетворяющей некоторым условиям оптимальности.

Варианты точечной оценки информационного множества

- центр интервального бруса, который является минимальной по включению внешней оценкой информационного множества,
- центр Оскорбина,
- чебышёвский центр,
- центр тяжести,
- точка максимума совместности (аргумент максимума распознающего функционала, который является точкой максимума совместности соответствующей интервальной системы уравнений).

Пример обработки накрывающей выборки.

Пример обработки накрывающей выборки

Пример иллюстрирует практическое применение методики главы «Задача восстановления зависимостей» книги «Обработка и анализ данных с интервальной неопределённостью» [1].

Технологически изложение следует канве, представленной в виде блокнота на ресурсе С.Жилина [4].

Установка

Ермаков Н.В., Баженов А.Н., Смирнов А.Н., Толстяков С.Ю. Стенд для испытаний шаговых двигателей. Приборы и техника эксперимента. 2023. – №1. – С. 151-152.

Данные

<https://drive.google.com/drive/folders/11haRRx4ZxVbym9LodfX-t1yuWAZa9RCo>

Набор данных.

При измерении параметров шагового двигателя была получена зависимость положения вала от времени.

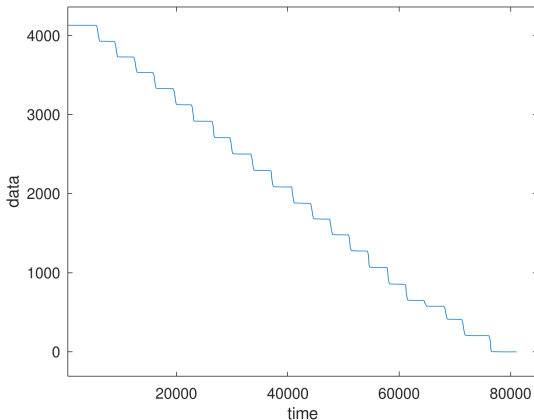


Рис.: Положение вала от времени. Данные энкодера углового перемещения.

Гистограмма данных.

На Рис. 11 горизонтальные участки соответствуют устойчивым положениям вала, а вертикальные — его повороту. Для выделения устойчивых положений, построим гистограмму

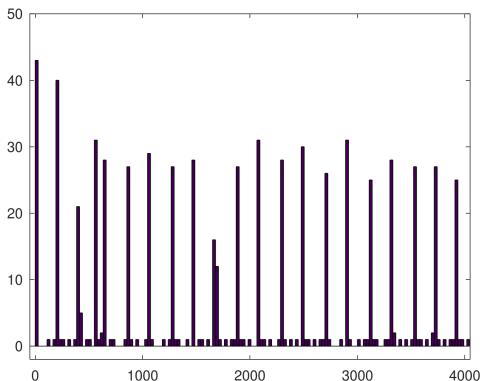


Рис.: Гистограмма положений вала двигателя.

На основании данных гистограммы Рис. 12 можно выделить устойчивые положения как те, в которых двигатель находился больше какого-то времени. Таким образом приходим к зависимости положения вала от номера шага.

Рис. 13 подобен Рис. 11 с заменой горизонтальных участков данных на одиночные значения. Сдвинуто начало отсчета энкодера, чтобы работать с более удобными для визуальной оценки числами. Также для удобства график показан возрастающим по коду энкодера.

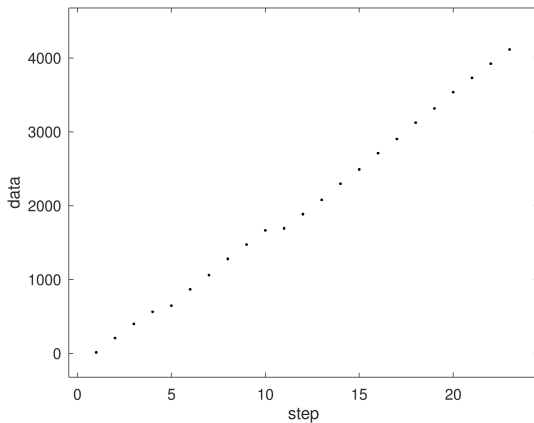


Рис.: Зависимость положения вала от номера шага.

Рабочая выборка.

Для удобства восприятия, выберем 10 значений замеров из числа данных, представленных на Рис. 13. Конкретно выбрано 10 первых нечётных значений для статических положений вала двигателя.

Номер измерения	Данные энкодера
1	399
2	646
3	1059
4	1472
5	1692
6	2078
7	2491
8	2904
9	3316
10	3729

Таблица: Частичная выборка данных.

Точечная оценка параметров регрессии.

Данные энкодера выдаются в виде целых десятичных значений, так что неопределённость представления — младший десятичный разряд. Реально погрешность, как мы увидим, существенно выше, и включает много факторов, о части которых неизвестно ничего.

В качестве первого подхода к проблеме, проведем точечную оценку параметров регрессии. Пусть модель задаётся в классе линейных функций

$$y = \beta_1 + \beta_2 x, \quad (7)$$

x — номер измерения в выборке Табл. 1,

y — угол поворота вала двигателя.

Точечная оценка параметров регрессии.

Для согласования с данными поставим задачу оптимизации и решим её методами линейного программирования [1].

$$\text{mid } \mathbf{y}_i - w_i \cdot \text{rad } \mathbf{y}_i \leq X\beta \leq \text{mid } \mathbf{y}_i + w_i \cdot \text{rad } \mathbf{y}_i, \quad i = 1, m, \quad (8)$$

$$\sum_{i=1}^m w_i \longrightarrow \min \quad (9)$$

$$w_i \geq 0, \quad i = 1, m, \quad (10)$$

$$w, \beta = ? \quad (11)$$

Здесь X — матрица $m \times 2$, в первом столбце которой элементы, равные 1, во втором — значения x_i .

В качестве значений середины и радиуса возьмём $\text{mid } \mathbf{y}_i = y_i$ и $\text{rad } \mathbf{y}_i = 1$.

Уравнение регрессионной прямой получилось

$$y = -50.2 + 369.4 \cdot x.$$

Точечная оценка параметров регрессии.

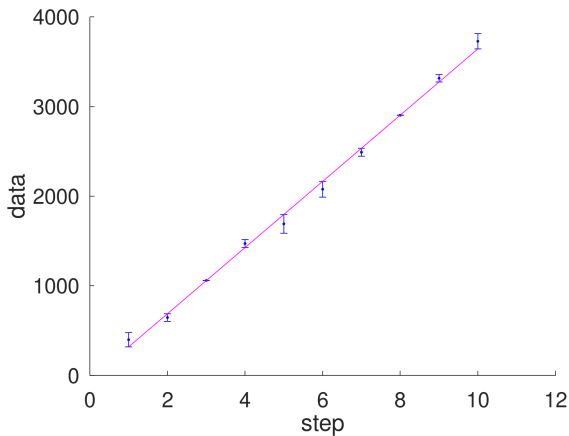


Рис.: Регрессия с оценкой по норме L_1 .

Вектор весов достижения совместности.

Вектор весов w радиусов отдельных замеров изображен на Рис. 15. Вместе с Рис. 14, высокая неоднородность значений w свидетельствует о разной по величине степени отклонении данных от регрессионной прямой на разных участках оси абсцисс.

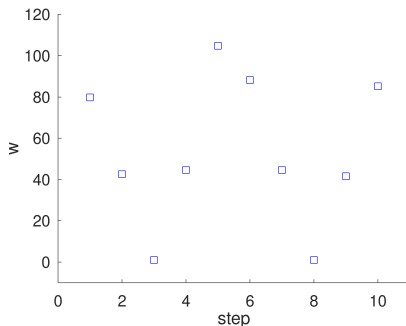


Рис.: Значения весов в задаче оптимизации.

Модель погрешности данных.

Приступим к интервальной оценке параметров регрессии. Ясно, при достаточно высокой погрешности данных выборка станет *накрывающей* или, по крайней мере совместной, согласно терминологии [1].

Для этого необходимо приписать данным какие-то погрешности. Значения компонент вектора w несут индивидуальную информацию о каждом измерении. Такая информация обладает высокой степенью избыточности, и желательно её заменить на более экономное представление.

Как видно из Рис. 15, имеет смысл в качестве первой оценки реалистичной погрешности данных взять близкой к максимальному значению w . Итак, пусть значение

$$\text{rad } \mathbf{y}_i := \varepsilon = \max_i w_i \simeq 150.$$

Диаграмма рассеяния данных.

Приведём диаграмму рассеяния данных для конкретного $\varepsilon = 150$ — Рис. 16.

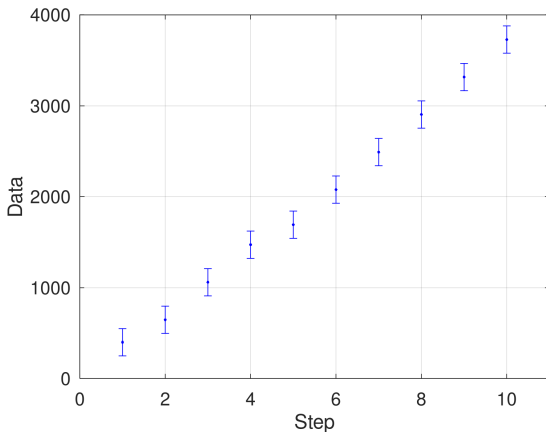
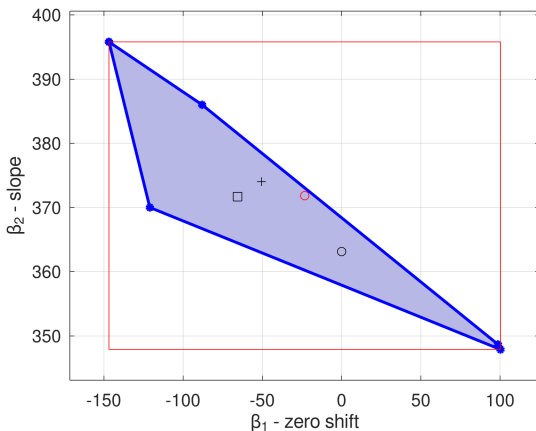


Рис.: Диаграмма рассеяния, погрешность данных $\varepsilon = 150$.

Информационное множество параметров I .

Определим теперь интервальные параметры регрессии по методике [4]. На Рис. 17 приведено информационное множество сдвигов и наклонов регрессионной прямой. Оно ограничено многоугольником и дано заливкой.



Информационное множество параметров I .

Также на Рис. 17 приведены различные точечные оценки.

Они достигнуты вычислением

- максимальной диагонали,
- центра тяжести,
- методом наименьших квадратов,
- точечной регрессией.

Для заданного значения погрешности данных все точечные оценки содержатся в информационном множестве.

Коридор совместности γ .

На Рис. 18 приведены диаграмма рассеяния данных и коридор совместности параметров модели регрессии для заданной погрешности данных.

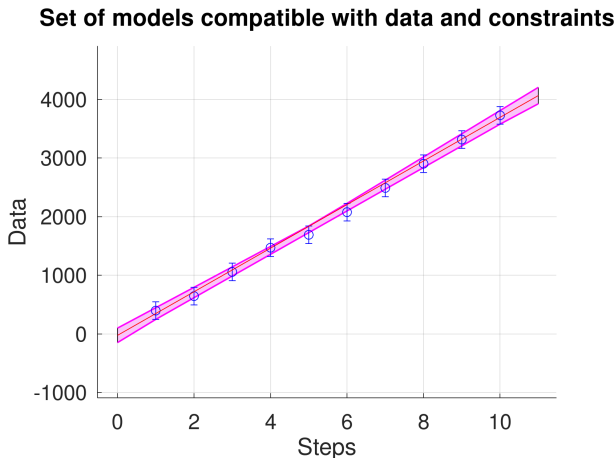


Рис. 18. Диаграмма рассеяния и коридор совместности параметров модели регрессии для заданной погрешности данных.

Коридор совместности γ .

Также дана прямая регрессии по параметрам, соответствующим центру тяжести множества, показанного на Рис. 17.

Для значения независимой переменной, равному 5, эта прямая касается границ коридора совместности.

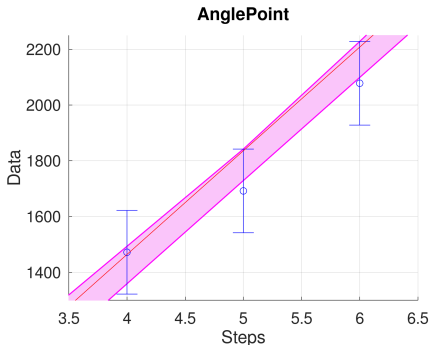


Рис.: «излом» множества γ .

Прогноз значений выходной переменной.

Важнейшим назначением регрессионной модели является предсказание значений выходной переменной для заданных значений входной.

С помощью построенной модели — Рис. 18

$$y(x) = [-146, 100] + [348, 396] \cdot x \quad (12)$$

можно получить прогнозные значения выходной переменной в точках эксперимента.

Прогноз значений выходной переменной.

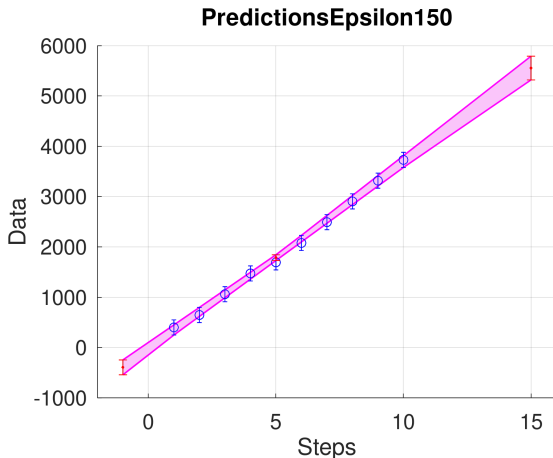


Рис.: Прогноз значений внутри и вне интервала имеющих данных, погрешность данных $\varepsilon = 150$.

Прогноз значений выходной переменной.

Ценность модели заключается в возможности её употребления для предсказания выходной переменной в точках, где измерения не производились. Для иллюстрации приведём прогнозы в одной точке внутри диапазона $x = 7$ и двух точках за его границами $x = -1, 13$. Результаты расчётов представлены в Табл. 2.

i	x_i	mid y	rad y_i	\underline{y}_i	\overline{y}_i
1	-1	-395.1	147.5	-542.6	-247.6
2	5	1785.5	56.5	1729	1842
3	15	5554.3	235.9	5318.4	5790.2

Таблица: Прогноз измерений по модели (12).

Прогноз значений выходной переменной.

Погрешность прогноза для «внутренней» точки $x = 5$ составляет $\simeq 56.5$ кодов энкодера и меньше назначенной погрешности 150.

При выборе точек прогноза со значениями -1 и 15 за пределами диапазона данных, даёт соответственно погрешность прогноза $\simeq 147$ и $\simeq 235$.

Чем более удалена точка прогноза от области данных, тем больше предсказываемая погрешность.

Уточнение модели погрешности данных.

Итак, при значении погрешности данных, равной $\varepsilon = 150$, получены согласованные оценки параметров линейной модели данных (12).

Напомним, что величина ε выбрана «с запасом» из соображений обеспечения заведомого согласования данных и линейной модели.

Посмотрим, что произойдёт при попытке уменьшить эту неопределённость. Пусть $\varepsilon = 100$.

Определим интервальные параметры регрессии. На Рис. 21 приведено новое информационное множество сдвигов и наклонов регрессионной прямой.

Информационное множество.

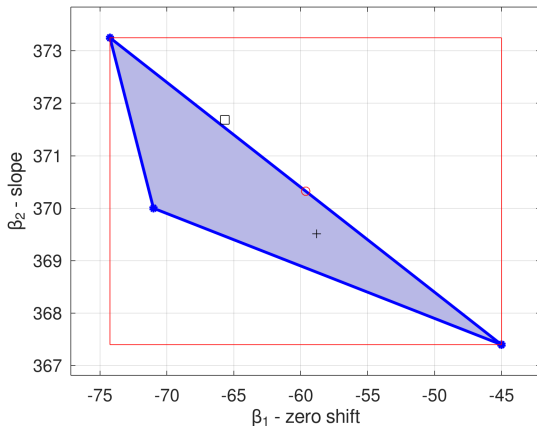


Рис.: Информационное множество, погрешность данных $\varepsilon = 100$.

Множество параметров линейной модели.

Множество параметров линейной модели на Рис. 21 существенно меньше аналогичного множества Рис. 17. Конкретные значения ширины параметров β приведены в Табл. 3.

ε	wid β_1	wid β_2
100	$\simeq 29$	$\simeq 4$
150	$\simeq 250$	$\simeq 46$

Таблица: Размеры множества параметров линейной модели.

Таким образом, информационное множество очень уменьшилось в размерах: примерно на десятичный порядок по каждой компоненте.

Согласование становится в таких условиях весьма проблематичным. В частности, оценка точечных параметров модели методом наименьших квадратов (черный квадратик на Рис. 21) находится за пределами I .

Уменьшение информационного множества приводит к сужению коридора совместности параметров модели. На Рис. 22 приведены диаграмма рассеяния данных и коридор совместности параметров модели регрессии \mathcal{X} для заданной погрешности данных.

Коридор совместности.

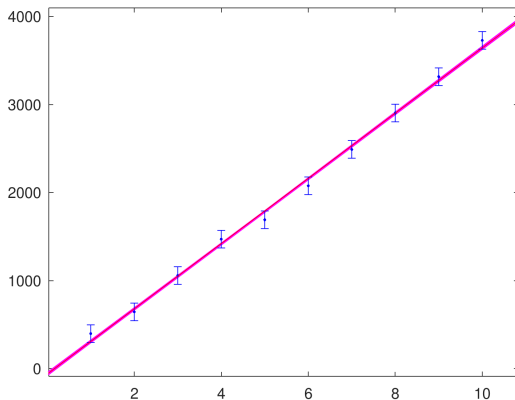


Рис.: Диаграмма рассеяния и коридор совместности \mathcal{Y} , погрешность данных $\varepsilon = 100$.

Коридор совместности.

Коридор совместности \mathcal{I} представляет собой узкую полосу, проходящую через крайние значения нескольких брусков.

Именно, коридор совместности касается вершин брусков

- \underline{y}_1 ,
- $\overline{y}_5, \overline{y}_6$,
- \underline{y}_{10} .

Как уже было замечено ранее, в середине графика имеется «излом».

Дальнейшее уменьшение ε приводит к пустоте множества параметров. Выборка становится *ненакрывающей*.

Граничными называют измерения, определяющие какой-либо фрагмент границы информационного множества. Очевидно, это свойство имеет смысл рассматривать для наблюдений, принадлежащих выборке S_n , по которой сконструирована модель и её информационное множество $\Omega(S_n)$.

Подмножество всех граничных наблюдений в S_n играет особую роль, поскольку оно является

минимальной подвыборкой, полностью определяющей модель.

Удаление неграничных наблюдений из выборки не изменяет модель.


Пример обработки накрывающей выборки — заключение.


В приведённом примере была продемонстрирована технология обработки выборки с *неизвестной заранее погрешностью данных*.


Выбором модели погрешностей выборка была сделана *накрывающей*.


Далее было показано, что при занижении погрешности данных происходит уменьшение информационного множества вплоть до его пустоты.

Пример обработки ненакрывающей выборки.

-  А.Н. БАЖЕНОВ, С.И. Жилин, С.И. Кумков, С.П. ШАРЫЙ. Обработка и анализ данных с интервальной неопределённостью. (готовится к изданию). с.300+.

-  А.Н. БАЖЕНОВ. Введение в анализ данных с интервальной неопределённостью. 2023.
<https://elib.spbstu.ru/dl/2/id22-247.pdf/info>

-  С.П. Шарый. Конечномерный интервальный анализ. — Новосибирск: XYZ, 2022. — Электронная книга, доступная на <http://interval.ict.nsc.ru/Library/InteBooks/SharyBook.pdf>

-  С.И.Жилин. Примеры анализа интервальных данных в Octave
<https://github.com/szhilin/octave-interval-examples>