Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Высшая школа прикладной математики и вычислительной физики

Отчёт по курсовой работе
на тему:
Восстановление зависимостей

Студент группы 5040102/00201: Курносов Д.А.
Преподаватель: Баженов А.Н.

Санкт-Петербург
2021 г.

Содержание

1 Постановка задачи

Построить для выбранного массива реальных данных соотвествующую модель регрессии и определить её параметры. Обработать данные на предмет выбросов. Построить прогноз за пределами экспериментальныъ данных.

2 Исходные данные

Основной целью построения интервальной регрессионной модели является восстановление функциональных зависимостей на основе конкретных наблюдений. В общем виде постановка задачи выглядит следующим образом:

$$y = f(x, \beta) - \text{некоторая функция}, \tag{1}$$

Здесь, $x = (x_1, ..., x_m)$ - вектор независимых переменных, $\beta = (\beta_1, ..., \beta_l)$ - вектор параметров функции. Необходимо для заданных значений x, y найти коэффициенты $\beta_1, ..., \beta_l$, удовлетворяющие условию (1).

На практике, значения x, y являются результатами измерений и предполагают неточность. Предполагая, что неточность ограничена, мы воспользуемся интервальным представлением, считая, что истинное значение измерения x_i лежит в интервале $\mathbf{x_i}$. Аналогичные размышления применяются к значениям y.

Исходные данные представляют собой выборку, состоящую из 6 измерений, отображающих зависимость мощности ДВС от количества оборотов. Будем считать что измерения мощности для фиксированных значений оборотов имеют относительную погрешность $\epsilon = 0.05$. Таким образом, исходные данные могут быть представленны в виде таблицы:

i	x	y$^- = y - \epsilon$	y	y$^+ = y + \epsilon$
1	2500	83.841	88.254	92.667
2	3000	99.26	104.485	109.71
3	3500	126.24	132.889	139.54
4	4000	163.82	172.451	181.08
5	4500	180.18	189.671	199.16
6	5000	198.52	208.971	219.42

Таблица 1: Исходные данные

Здесь i — номер измерения, y_i — значение мощности ДВС при фиксированном в i-м опыте значении оборотов, $y_i^- = y_i - \varepsilon_i$ — нижняя граница интервального измерения

$\mathbf{y}_i$, $y_i^+ = y_i + \varepsilon_i$ — верхняя граница интервального измерения $\mathbf{y}_i$.
Отобразим исходные данные графически:

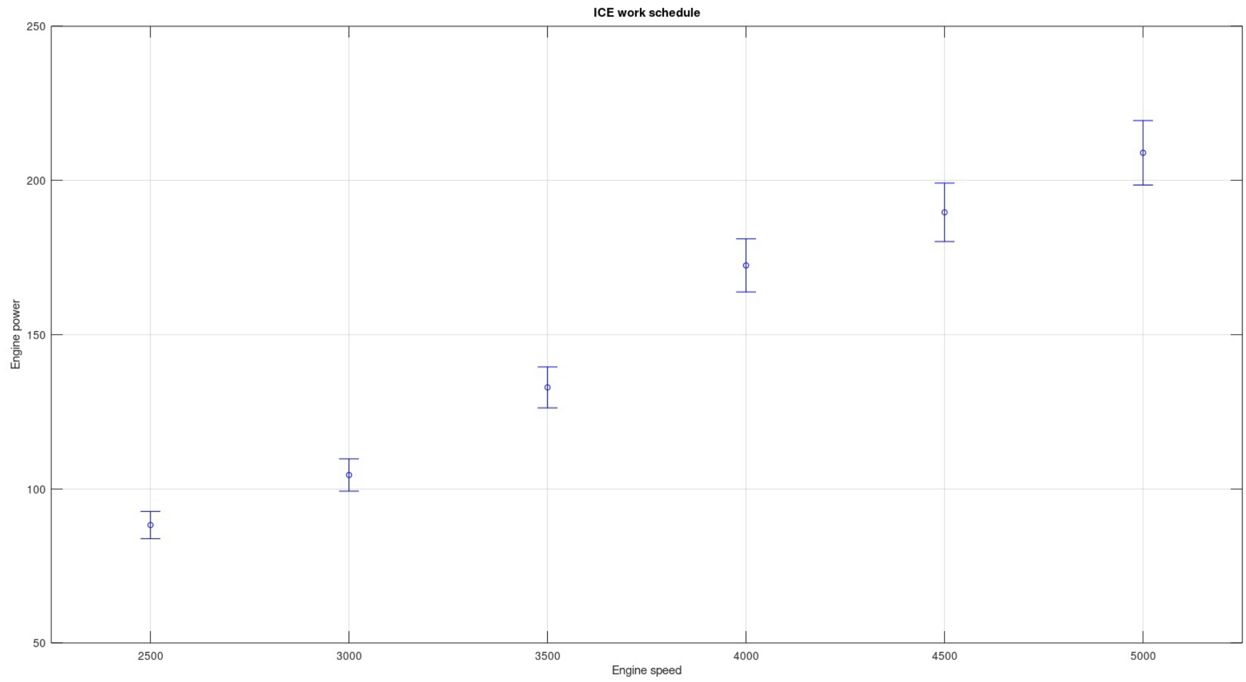


Рис. 1: Исходные данные

В рамках данной задачи мы будем искать модель в классе линейных функций вида:

$$y = \beta_1 + \beta_2 x \tag{2}$$

3 Ход работы

3.1 Первичные данные и информационное множество

Построим линейную модель с помощнью МНК для точечных значений измерений. Полученные значение коэффициентов $\beta_1 = -43.126, \beta_2 = 0.0513$. График модели в интересующих нас ограничениях выглядит следующим образом:

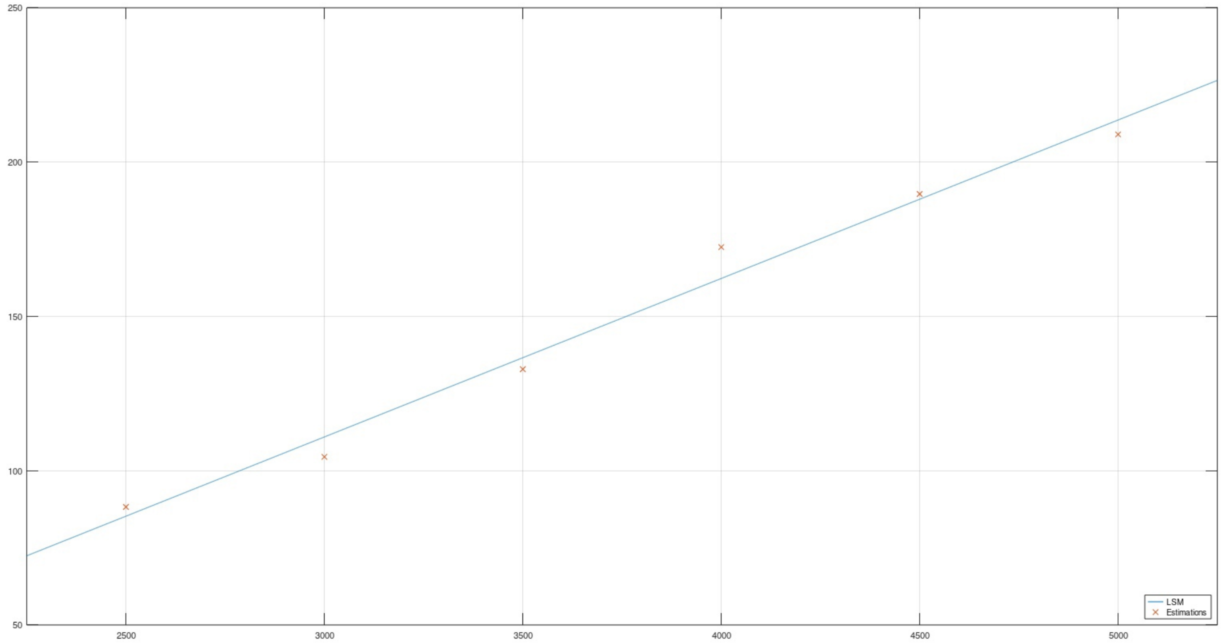


Рис. 2: Модель МНК

Вернемся к интервальным данным. Пытаясь построить модель в классе линейных функций для данных измерений мы получаем следующую систему неравенств:

$$83.841 \leq \beta_1 + 2500\beta_2 \leq 92.667,$$

$$99.26 \leq \beta_1 + 3000\beta_2 \leq 109.71,$$

$$126.24 \leq \beta_1 + 3500\beta_2 \leq 139.54,$$

$$163.82 \leq \beta_1 + 4000\beta_2 \leq 181.08,$$

$$180.18 \leq \beta_1 + 4500\beta_2 \leq 199.16,$$

$$198.52 \leq \beta_1 + 5000\beta_2 \leq 216.42,$$

Множество решений этой системы является информационным множеством. Но при попытке построить такое множество для нашей задачи, мы обнаружим, что оно пусто.

Это может происходить из-за неверно выбранного значения погрешности, которое не совпадает с реальным значением. Для решения этой проблемы решим следующую задачу оптимизации:

$$\begin{gathered} mid\boldsymbol{y_i} - w_i \cdot rad\boldsymbol{y_i} \leq (\boldsymbol{X\beta})_i \leq mid\boldsymbol{y_i} + w_i \cdot rad\boldsymbol{y_i}, \\ \sum_{i=1}^{m} w_i \to min, \\ w_i \geq 0, \ \ i = 1, ..., m. \end{gathered} \tag{3}$$

Получаем следующие оценки: $w = [1, 1.15, 1, 1, 1, 1], \beta = [-49.4, 0.05]$. Преобразуя нашу выборку в соответствии с данными оценками, мы получаем новую выборку. Её графическое представление имеет вид:

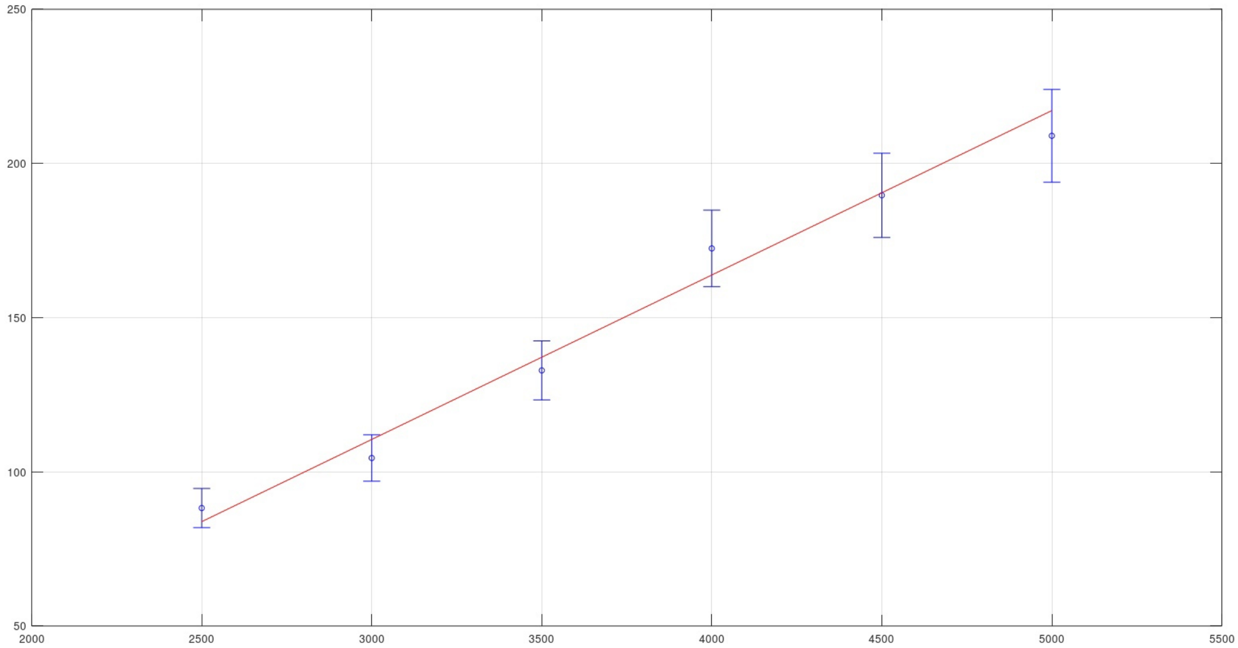


Рис. 3: Преобразованные данные

С учетом того, что максимальный коэффициент масштабирования для всех измерений равен 1.15, можно утверждать, что выборка не содержит выбросов. Построим информационное множество для преобразованной выборки:

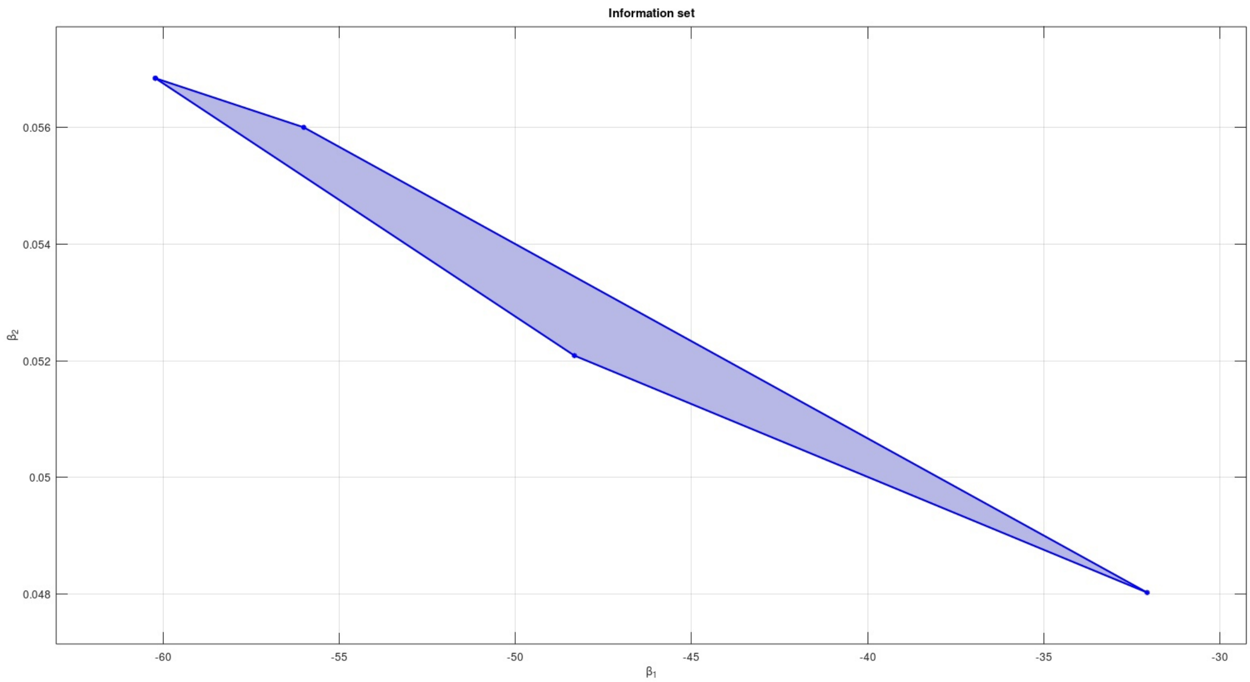


Рис. 4: Информационное множество

Расстояние между наиболее удалёнными вершинами (диаметр этого многогранника $\rho(B)$), может служить мерой неопределённости задачи.
Оценим параметры модели регрессии, используя данное информационное множество.
Внешняя интервальная оценка параметра определяется минимальным и максимальным значениями, которых может достигать значение параметра в информационном множестве.

$$\beta_1 = [-60.21, -32.06],$$

$$\beta_2 = [0.048, 0.0568].$$

В совокупности интервальные оценки параметров задают брус, описанный вокруг информационного множества и именуемый внешней интервальной оболочкой информационного множества.
Точечной оценкой параметров может быть любая из точек информационного множества. В качестве некоторой центральной оценки могут быть выбраны следующие варианты:
Центр наибольшей диагонали информационного множества

$$\hat{\beta}_{\text{maxdiag}} = 0.5(b_1 + b_2) = [-46.14, 0.0524],$$

Центр тяжести информационного множества

$$\hat{\beta}_{\text{gravity}} = \frac{1}{4}\sum_{i=1}^{4} b_i = [-49.15, 0.0532].$$

Графически данные оценки выглядят следующим образом:

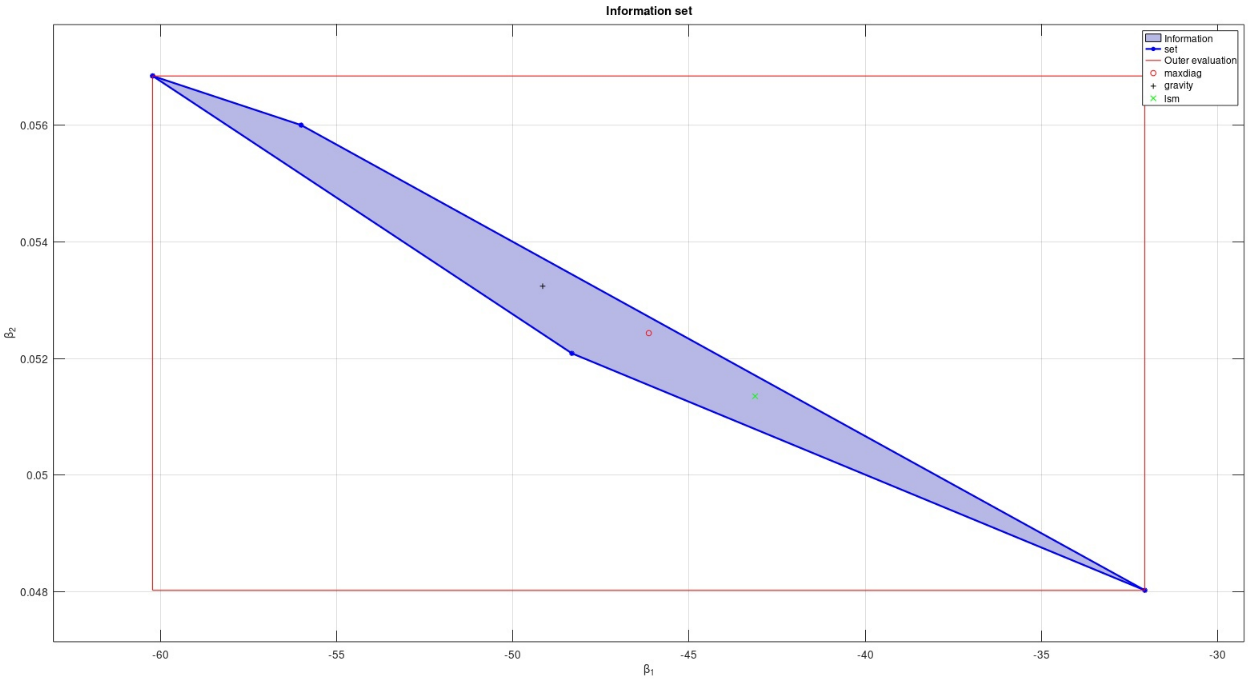


Рис. 5: Оценки параметров

3.2 Коридор совместных зависимостей

Определив параметры функциональной зависимости, мы можем предсказать значения зависимости в других интересующих нас точках области определения. Однако такое предсказание будет осуществляться с некоторой погрешностью, вызванной неопределённостями данных, неоднозначностью самой процедуры восстановления и т. п.

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, 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. Такое множество называется коридором совместных зависимостей.

Графическое представление коридора совместных зависимостей для нашей модели, с параметрами оцененными как центр наибольшей диагонали информационного множества, выглядит следующим образом:

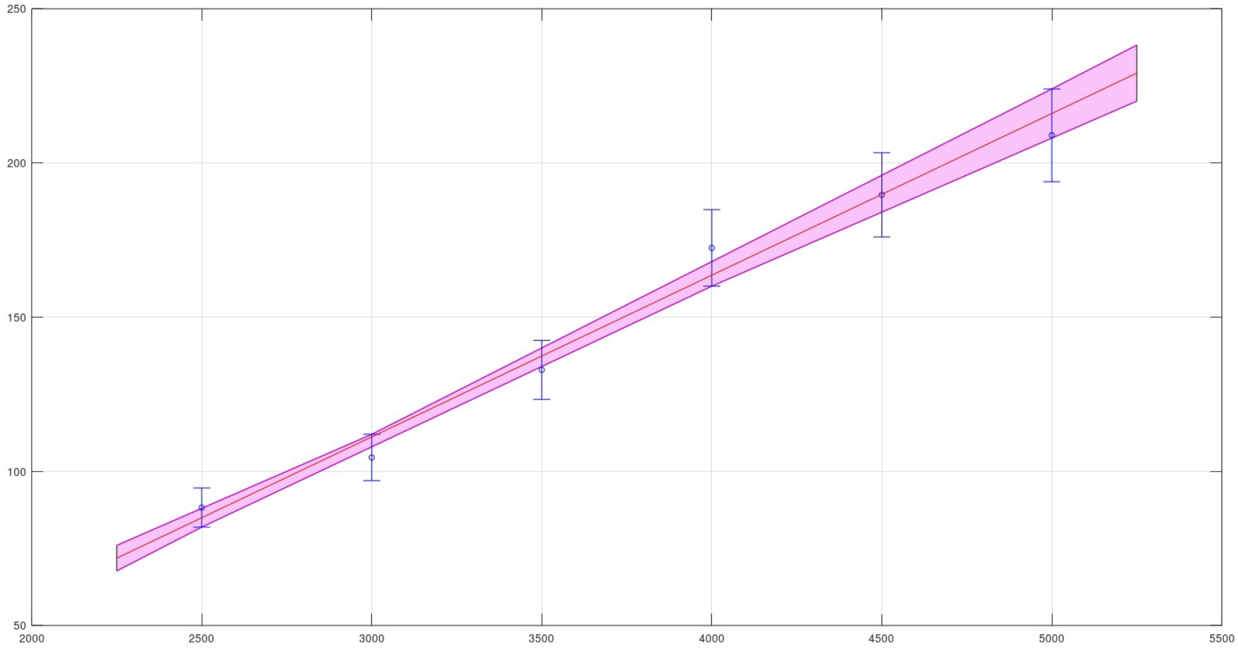


Рис. 6: Коридор совместных зависимостей

3.3 Прогноз значений выходной переменной

Полученные ранее интервальные оценки параметров позволяют построить модель вида:

$$\mathbf{y}(x) = [-53.49, -45.53] + [0.052, 0.0546]x$$

Проверим корректность полученной модели с помощью точек, являющихся первичными данными. Предсказанные значения y имеют вид:

$$y_1 = [81.899, 87.996]$$

$$y_2 = [107.944, 112.008]$$

$$y_3 = [133.989, 140.011]$$

$$y_4 = [160.034, 168.013]$$

$$y_5 = [184.047, 196.015]$$

$$y_5 = [208.059, 224.018]$$

Видно, что наиболее удачным оказалось 5 измерение, точечное значение коротого имеет погрешность относительно исходного значения менее 1 процента. Погрешность 2 измеренения имеет самое высокое значение в 5.2 процента. Можно сказать, что полученая модель достаточно точно описывает нашу выборку.
Воспользуемся теперь построенной моделью для прогнозирования значений y для значений x не входящих в выборку, а именно $[2000, 3250, 3750, 4250, 5500]$ Коридор зависимости для данных точек имеет вид:

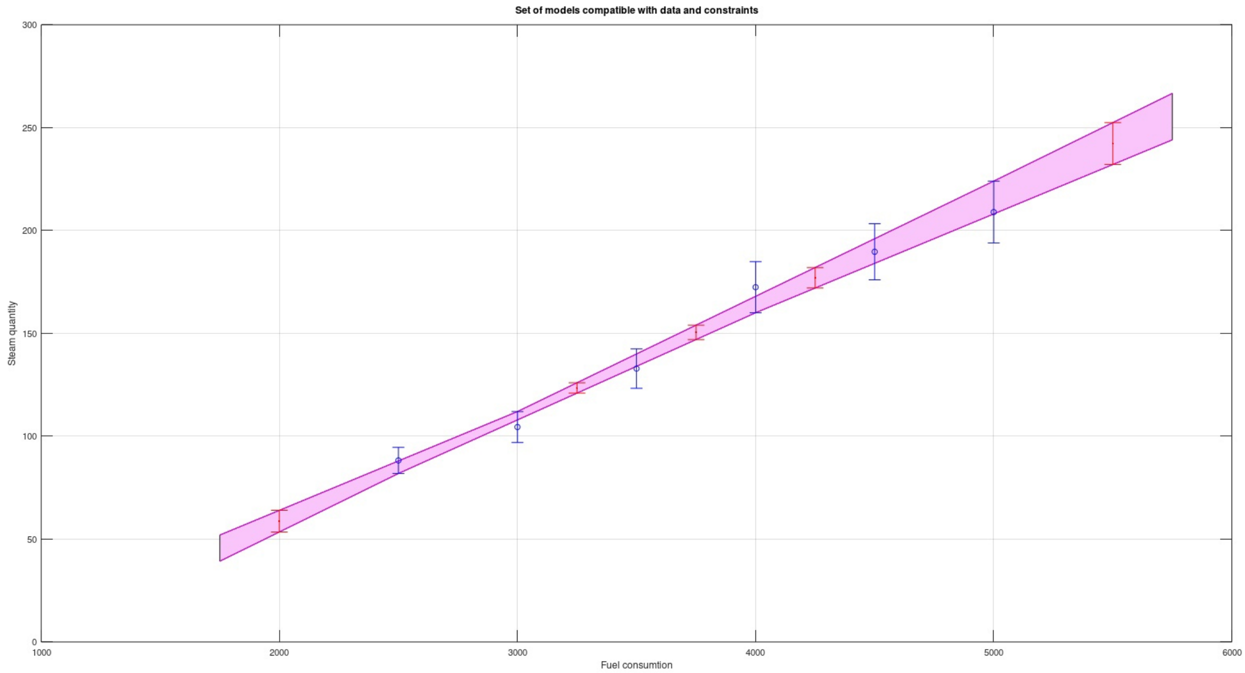
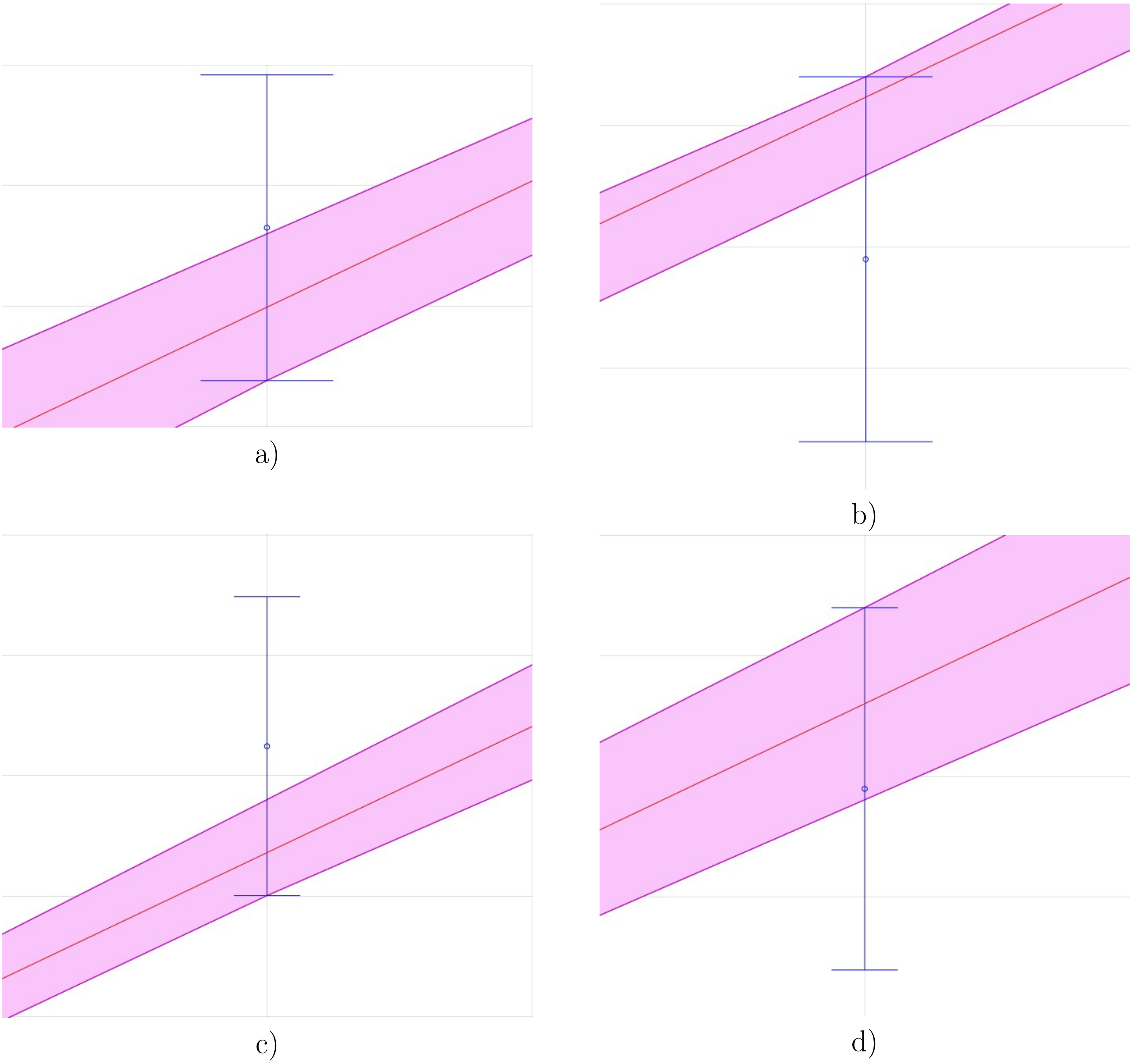


Рис. 7: Коридор совместных зависимостей для предсказанных значений

На графике видно, что величина неопределённости прогнозов растёт по мере удаления от области, в которой производились исходные измерения. Это обусловлено видом коридора зависимостей, расширяющимся за пределами области измерений, и согласуется со здравым смыслом.

3.4 Граничные точки множества совместности

Граничными называют измерения, определяющие какой-либо фрагмент границы множества. Очевидно, это свойство имеет смысл рассматривать для наблюдений, принадлежащих выборке, по которой была построена модель. Граничные измерения задают минимальную подвыборку, определяющую модель. На Рис. 6 подходящими точками выглядят $1, 2, 4$ и 6. Рассматривая каждую из них отдельно, мы можем в этом убедиться.



a) b) c) d)

Рис. 8: Граничные точки: a) Первое измерение, b) Второе измерение, c) Четвертое измерение, d) Шестое измерение.

Таким образом, грацичными точками являются $1, 2, 4$ и 6.

4 Заключение

В результате выполнения данной работы нам удалось для заданной выборки измерений, используя методы линейного программирования получить вектор весов достижения совместности, благодаря которому мы оценили величину реальной погрешности. Используя данную оценку и скорректировав выборку, мы посмтроили регрессионную модель, которая достаточно точно описывает полученные данные и позволяет сделать прогноз значений, выходящих за пределы выборки.

5 Литература

- Вощинин А.П., Сотиров Г.Р. "Оптимизация в условиях неопределенности."М.: Изд-во МЭИ; София: Техника, 1989. – 224 с.

- А.Н.Баженов, С.И. Жилин, С.И. Кумков, С.П.Шарый "Обработка и анализ данных с интервальной неопределённостью. 2021.

- Sergei Zhilin GitHub URL: https://github.com/szhilin/octave-interval-examples

6 Приложение

- Ссылка на GitHub с реализацией: https://github.com/ExpressFromSiberia/StochasticModels