МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

КАФЕДРА МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ И АНАЛИЗА ДАННЫХ

Отчет
о прохождении преддипломной практики

Румянцева Андрея Кирилловича
студента 4 курса, специальность
"прикладная математика"

Руководитель практики:
зав. кафедрой ММАД,
канд. физ.-мат. наук, доцент
Бодягин Игорь Александрович

Минск, 2019

1 Задание на практику

- Провести аналитический обзор литературы методов статанализа данных при наличии классифицированных данных с искажениями.
- Реализовать альтернативные встречаемые в литературе методы статистического анализа данных при наличии классифицированных данных с искажениями.
- Провести сравнительный анализ реализованного в ходе курсового проекта метода с альтернативными.
- Обобщить все реализованные методы с линейной на полиномиальную регрессию.
- Подготовить отчет по преддипломной практике.

Содержание

ВВЕДЕНИЕ

Целью преддипломной практики было продолжение исследования и улучшение оценок, построенных в курсовом проекте. Темой курсового проекта было *"Статистическое оценивание параметров линейной регрессии с выбросами при наличии группирования наблюдений"*.

Оценки были построенны с помощью максимизирования функции правдоподобия:

$$l(\beta, \sigma^2, \nu_0, \ldots, \nu_{k-1}) = \ln(\prod_{i=1}^{n} P(\mu_i = j)) = \tag{1}$$

$$= \sum_{i=1}^{n} \ln(P(\mu_i = j)), \tag{2}$$

где:

$$P(\mu_i = j) = P(y_i \in \nu_{\mu_i}), \tag{3}$$

y_i - значения функции регрессии, а $\nu_0, \ldots, \nu_{k-1}$ - номера полуинтервалов, разбивающих множество значений функции регрессии:

$$(-\infty, a_1] \bigcup (a_1, a_2] \bigcup \cdots \bigcup (a_{k-1}, +\infty) = \mathcal{R} \tag{4}$$

μ_i номер полуинтервала, в который он попал y_i.

$$\mu_i = j, \text{если } y_i \text{ отнесли к полуинтервалу } \nu_j. \tag{5}$$

Задача максимизирования решала с помощью решения нелинейной системы уравнений:

$$\frac{\delta l}{\delta \beta} = 0, \tag{6}$$

где

$$\frac{\delta l}{\delta \beta} = \frac{\delta \sum_{i=1}^{n} \ln(P(\mu_i = j))}{\delta \beta} = \frac{\delta \sum_{i=1}^{n} \ln P(y_i \in \nu_{\mu_i})}{\delta \beta} =$$
$$= \frac{\delta \sum_{i=1}^{n} \ln(\frac{1}{2}(\text{erf}(\frac{a_{\mu_i+1} - f(x_i,\beta)}{\sqrt{2}\sigma}) - \text{erf}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma})))}{\delta \beta} =$$
$$= \sum_{i=1}^{n} \Big((1 - (\delta_{\mu_i 0} + \delta_{\mu_i k-1})) \frac{(\text{erf'}(\frac{a_{\mu_i+1} - f(x_i,\beta)}{\sqrt{2}\sigma}) - \text{erf'}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma}))}{(\text{erf}(\frac{a_{\mu_i+1} - f(x_i,\beta)}{\sqrt{2}\sigma}) - \text{erf}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma}))} +$$
$$+ (\delta_{\mu_i 0} + \delta_{\mu_i k-1}) \frac{\text{erf'}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma})}{(1 + \text{erf}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma}))} \Big) (-1) \frac{\delta f(x_i, \beta)}{\delta \beta}) = \tag{7}$$

$$= -\sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_{i1} \\ \cdots \\ x_{in} \end{pmatrix} \times \Big((1-(\delta_{\mu_i 0}+\delta_{\mu_i k-1})) \frac{(\text{erf'}(\frac{a_{\mu_i+1}-f(x_i,\beta)}{\sqrt{2}\sigma}) - \text{erf'}(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma}))}{(\text{erf}(\frac{a_{\mu_i+1}-f(x_i,\beta)}{\sqrt{2}\sigma}) - \text{erf}(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma}))} +$$

$$+ (\delta_{\mu_i 0} + \delta_{\mu_i k-1}) \frac{\text{erf'}(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma})}{(1 + \text{erf}(\frac{a_{\mu_i}-f(x_i,\beta)}{\sqrt{2}\sigma}))} \Big).$$

2 Изучение материала

В ходе выполнения преддипломной практики были изучены следующие источники:

В источниках был встречен *метод наименьших квадратов по центрам интервалов*. Метод заключается в следующем: пусть имеется μ_i - номер полуинтервала, в который попало очередное наблюдение y_i. Ему соответствует полуинтервал ν_{μ_i} (см(4)), т.е. полуинтервал:

$$(a_{\nu_{\mu_i}}, a_{\nu_{\mu_i}+1}], \tag{8}$$

(считаем что $a_1 < y_i < a_{k-1}, i = \overline{1,n}$).

Найдем центральную точку этого интервала, т.е. точку

$$\check{y}_i = \frac{a_{\nu_{\mu_i}} + a_{\nu_{\mu_i}+1}}{2} \tag{9}$$

Построим для всех значений функции регрессии y_i значения $\check{y}_i$. Будем использовать в качестве значений функции регрессии полученные значений, а в качестве регрессоров x_i и построим МНК оценки параметров β.

3 Реализация оценка

Описанные оценки были построены путем наследования от исходных оценок и переопределения соответствуюего метода fit().

```
class ApproximationGEMModelNaive(ApproximationGEMModelRedesigned):
    def fit(self):
        self.classify()

        def ex_generator(mu_data):
            for i in range(0, self.endogen.size):
                if mu_data[i] is None:
                    continue
                a_mu_i_plus_1 = mu_data[i] * Defines.INTERVAL_LENGTH
                a_mu_i = mu_data[i] * Defines.INTERVAL_LENGTH - Defines.INTERVAL_LENGTH
                yield (a_mu_i_plus_1 + a_mu_i) / 2

        naive_ex_data_positive = np.fromiter(ex_generator(self._np_freq_positive), float)
        naive_ex_data_negative = np.fromiter(ex_generator(self._np_freq_negative), float)

        naive_ex_data_full = np.append(naive_ex_data_positive, naive_ex_data_negative)

        z, resid, rank, sigma = np.linalg.lstsq(self.exogen, naive_ex_data_full, rcond=None)
        return z
```

4 Компьютерные эксперименты

4.1 Параметры модели и оценок

Параметры программы	
Переменная	значение
Размер выборки N	1000
Доля выбросов $\widetilde{\varepsilon}$	0.8
Параметры регрессии β	$(90, 4)$
Регрессоры x_i	$\sim U(-5, 5)$
ε_i	$\sim N(0, 16)$
η_i	$\sim N(100, 100)$
Величина K из пункта 2.3 курсового проекта	10

,

4.2 Сравнительный анализ построенной оценки с альтернативной

Если сравнить вариации оценок построенныхх на рис.3 можно увидеть, что оценки, построенные по методу, предложенному на курсовом проекте, показывают лучшие результаты

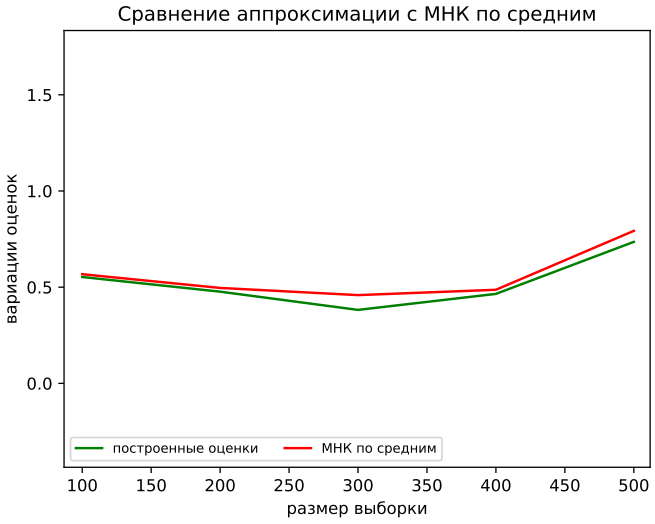

Рис. 1: Сравнение вариаций оценок

4.3 Дополнительные эксперименты

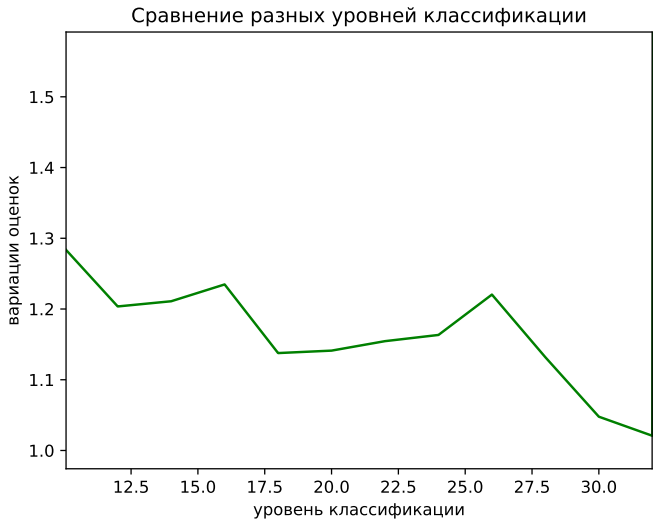

Рис. 2: Зависимость от K, упомянотого в пункте 2.3 курсового проекта

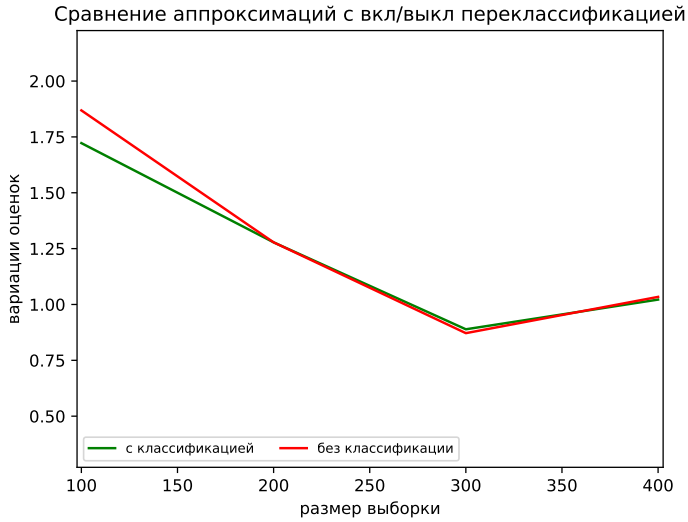

Рис. 3: Сравнение вариаций оценок когда используется и не используется переклассификация

4.4 Использование полиномиальной регрессии

Несложно заметить, что построенные в курсовом проекте оценки никак не зависят от регрессоров, поэтому можно моделировать полиномиальную регрессию и применить к ней описанный метод.