# Раннее прогнозирование достаточного объема выборки для обобщенной линейной модели.

***Бучнев В. С., Грабовой А. В., Гадаев Т. Т., Стрижов В. В.***

Исследуется проблема снижения затрат на сбор данных, необходимых для построения адекватной модели. Рассматриваются задачи линейной и логистической моделей. Для решения этих задач требуется, чтобы выборка содержала необходимое число объектов. Требуется предложить метод вычисления оптимального обьема данных, соблюдая при этом баланс между точностью модели и и трудозатратами при сборе данных. Предпочтительны те методы оценки объемы, которые позволяют строить адекватные модели по выборкам возможно меньшего объема.

**Ключевые слова**: *Обобщенная линейная модель, размер выборки.*

## 1 Введение

При планировании эксперимента требуется оценить минимальный объём выборки — число производимых измерений набора показателей или признаков, необходимый для построениие сформулированных условий.

Существует большое количество оценки размера выборки. Например, тест множителей Лагранжа, тест отношения правдоподобия и тест Вальда. В работах [1–3] на основе данных методов построена оценка оптимального размера выборки. Основной минус этих методов заключается в том, что статистики, используемые в критериях, имеют асимптотическое распределение и требуют большого обьема выборки.

Существуют также байесовские оценки обьема выборки: критерий средней апостериорной дисперсии, критерий среднего покрытия, критерий средней длины и метод максимизации полезности. Первые три метода требуют анализа некоторой функции эффективности от размера выборки. Используя некоторое решающее правило, по данной функции определяется достаточный объем выборки. Главный минус этих методов заключается в том, что они не позволяют построить аппроксимацию функции эффективности при большем объеме данных. Метод максимизации полезности максимизирует ожидание некоторой функции полезности по обьему выборки. Все эти методы опираются на апостериорное распределение, что требует достаточно большого обьема выборки.

Предлагается исследовать зависимость среднего значения логарифма правдоподобия от размера доступной выборки, а также его дисперсию. В данной работе предлагается использовать не сами функции эффективности, а их аппроксимации. Для этого предлагается использовать аппроксимацию ковариационной матрицы вектора параметров. После чего аппроксимировать данные две зависимости при помощи метода бутстреп. Для вычислительного эксперимента предлагается использовать классические выборки из UCI репозитория и синтетические данные.

## 2 Постановка задачи

Дана выборка размера m:

$$\mathfrak{D}_m = \{\mathbf{x}_i, y_i\}_{i=1}^m,$$

где $\mathbf{x}_i \in \mathbb{R}^n$ - вектор признаков, $y_i \in \mathbb{Y}$.

Предполагается, что выборка $\mathfrak{D}_m$ не противоречит гипотезе порождения данных.

Рассмотрим параметрическое семейство функций для аппроксимации неизвестного распределения $p(y|\mathbf{x}, \mathbf{w})$, где $\mathbf{w} \in \mathbb{W}$ - вектор параметров:

$$\mathfrak{F} = \left\{ f(y, \mathbf{x}, \mathbf{w}) | \mathbf{w} \in \mathbb{W}, \int_{y \in \mathbb{Y}, \mathbf{x} \in \mathbb{R}^n} f(y, \mathbf{x}, \mathbf{w}) dy d\mathbf{x} = 1 \right\}.$$

Для модели $f$ с вектором параметров $\mathbf{w}$ определим функцию правдоподобия и логарифмическую функцию правдоподобия выборки $\mathfrak{D}$:

$$L(\mathfrak{D}, \mathbf{w}) = \prod_{y, \mathbf{x} \in \mathfrak{D}} f(y, \mathbf{x}, \mathbf{w}), \quad l(\mathfrak{D}, \mathbf{w}) = \sum_{y, \mathbf{x} \in \mathfrak{D}} \log f(y, \mathbf{x}, \mathbf{w}),$$

где $f(y, \mathbf{x}, \mathbf{w})$ - аппроксимация плотности апостериорной вероятности выборки $\mathfrak{D}_{\mathcal{L}_m}$ при заданном векторе параметров $\mathbf{w}$.

Рассмотрим правдоподобие выборки $\mathfrak{D}_{\mathcal{L}_m}$:

$$L(\mathfrak{D}_{\mathcal{T}_m}, \mathfrak{D}_{\mathcal{L}_m}) = \prod_{y, \mathbf{x} \in \mathfrak{D}_{\mathcal{T}_m}} f(y, \mathbf{x}, \mathbf{w}).$$

Рассмотрим логарифм правдоподобия выборки $\mathfrak{D}_{\mathcal{L}_m}$:

$$l(\mathfrak{D}_{\mathcal{T}_m}, \mathbf{w}) = \sum_{y, \mathbf{x} \in \mathfrak{D}_{\mathcal{T}_m}} \log f(y, \mathbf{x}, \mathbf{w}).$$

Будем рассматривать ожидаемое значение функции $l$:

$$\tilde{l}(\mathfrak{D}) = \mathop{\mathsf{E}}_{y, \mathbf{x} \in \mathfrak{D}} l(\{y, \mathbf{x}\}, \mathbf{w}).$$

Рассмотрим ожидаемое значение логарифма правдоподобия по разным обучающим выборкам $\mathfrak{D}_{\mathcal{L}_m}$ размера $m^*$:

$$l(m^*) = \mathop{\mathsf{E}}_{\mathfrak{D}_{\mathcal{L}_m}} \tilde{l}(\mathfrak{D}_{\mathcal{L}_m}).$$

Будем считать, что объем выборки достаточный, если:

$$\forall m_1, m_2 > m^* \quad |l(m_1) - l(m_2)| < \delta,$$

где $\delta$ - достаточно малое пороговое значение.

Для оценки вектора параметров используется принцип максимума правдоподобия:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w} \in \mathbb{W}} L(\mathfrak{D}_{\mathcal{L}_m}, \mathbf{w}).$$

Для линейной регрессии $\mathbb{Y} = \mathbb{R}$, где $y$ представимо в виде:

$$y = \mathbf{x}^\top \mathbf{w} + \varepsilon,$$

где $\varepsilon \sim \mathcal{N}(0, 1)$.

Аппроксимация плотности апостериорной вероятности имеет вид:

$$f(y, \mathbf{x}, \mathbf{w}) = \mathcal{N}(y | \mathbf{x}^\top \mathbf{w}, 1).$$

Для логистической регрессии $\mathbb{Y} = \{0, 1\}$, где $y$ является бернуллиевской случайной величиной:

$$y \sim \mathcal{B}e(\theta),$$

где $\theta$ - неизвестный параметр распределения. Аппроксимация плотности апостериорной вероятности имеет вид:

$$f(y, \mathbf{x}, \mathbf{w}) = \mathcal{B}e(y|\theta), \quad \theta = \frac{1}{1 + exp(-\mathbf{x}^\top \mathbf{w})}.$$

# 3 Анализ эффективности модели

Методы байесовских оценок объема выборки основаны на ограничении некоторой выбранной характеристики модели. Для анализа эффективности вводится функция от объема выборки, увеличение значений которой интерпретируется как уменьшение эффективности модели. Объем выборки $m^*$ выбирается таким, при котором исследуемая функция не превышает некоторого порогового значения $\varepsilon$.

## 3.1 Критерий средней длины (ALC)

Пусть $A(\mathfrak{D}) \subset \mathbb{R}^n$ — множество значений параметров модели $\mathbf{w}$:

$$A(\mathfrak{D}) = \{\mathbf{w} : ||\mathbf{w} - \hat{\mathbf{w}}|| \leqslant r_m\}$$

такое, что

$$\mathsf{P}(A(\mathfrak{D})) = 1 - \alpha,$$

где $\alpha$ — некоторое малое значение.

Критерий средней длины выглядит следующим образом:

$$\forall m \geqslant m^* \ \mathsf{E}_{\mathfrak{D}_m} r_m \leqslant l,$$

где $r_m$ — радиус шара $A(\mathfrak{D}_m)$, $l$ — некоторое наперед заданное достаточно малое значение.

## 3.2 Критерий среднего покрытия (ACC)

Данный критерий также опирается на покрытие множества значений параметров модели $\mathbf{w}$.

Критерий среднего покрытия для определения $m^*$:

$$\forall m \geqslant m^* \ \mathsf{E}_{\mathfrak{D}_m} \mathsf{P}\{\mathbf{w} \in A(\mathfrak{D}_m)\} \geqslant 1 - \alpha,$$

где $\alpha$ — некоторое малое значение.

# 4 Описание алгоритма

## 4.1 Методы байесовских оценок объема выборки.

Для вычисления функции эффективности в методах ALC и ACC воспользуемся несмещеностью и состоятельностью оценки $\hat{\mathbf{w}}$:

$$\mathsf{E}\hat{\mathbf{w}} = \mathbf{m}, \quad \mathsf{D}\hat{\mathbf{w}} = \mathbf{I}^{-1}(\mathfrak{D}_m),$$

где $\mathbf{I}(\mathfrak{D}_m)$ — информационная матрица Фишера:

$$\mathbf{I}(\mathfrak{D}_m) = \mathsf{E}_{\mathbf{w}} \left( \frac{\partial L(\mathfrak{D}_m, \mathbf{w})}{\partial \mathbf{w}} \right)^2$$

Далее используется предположение о распределении оценки вектора параметров:

$$\hat{\mathbf{w}} \sim \mathcal{N}(\mathbf{m}, \mathbf{I}^{-1}(\mathfrak{D}_m)),$$

и с помощью сэмплирования вычисляется приближенное значение $r_m$.

### 4.2 Модификация алгоритма

В случае, когда $m \geqslant m^*$, воспользуемся свойством информационной матрицы Фишера:

$$\mathbf{I}(\mathfrak{D}_m) = m\ \mathbf{I}(\mathfrak{D}_1).$$

Получаем аппроксимацию матрицы Фишера для $m$ наблюдений:

$$\hat{\mathbf{I}}(\mathfrak{D}_m) = \frac{m}{m^*}\ \mathbf{I}(\mathfrak{D}_{m^*}).$$

Матрица $\hat{\mathbf{I}}(\mathfrak{D}_m)^{-1}$ является ковариационной матрицой оценки вектора параметров для $m$ наблюдений.

Таким образом, построена аппроксимация параметров распределения $\hat{\mathbf{w}}$ для $m$ наблюдений для вычисления приближенного значения функции эффективности:

$$\hat{\mathbf{w}} \sim \mathcal{N}(\mathbf{m}, \frac{m}{m^*}\ \mathbf{I}^{-1}(\mathfrak{D}_{m^*})).$$

## 5 Эксперимент на синтетической выборке

Рассматривается такая выборка $\mathbf{x}_1, ..., \mathbf{x}_n$, что компоненты вектора $x_i \in \mathbb{R}^n$ — реализации случайной величины из распределения $\mathcal{N}(0, 1)$,

$$\varepsilon_1, ..., \varepsilon_m \sim \mathcal{N}(0, 1).$$

Решается задача линейной регрессии с целевой переменной

$$y_i = \mathbf{w}^\top \mathbf{x}_i + \varepsilon_i$$

и функцией ошибки $MSE(\mathbf{x}, y, \mathbf{w}) = \frac{1}{m}\sum\limits_{i=1}^{m}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2$.

где $\mathbf{w}$ — реализация случайной величины из многомерного равномерного распределения $\mathcal{U}(\mathbf{0}, \mathbf{1})$.

Оценка параметра $\hat{\mathbf{w}}$ определяется следующим образом:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} MSE(\mathbf{x}, y, \mathbf{w}).$$

На графике 1 отображена зависимость значения функции ошибки от объема выборки. Видно, что при $m > 100$ значение функции ошибки почти не меняется, и увеличение объёма выборки не приводит к улучшению модели.

На графике 2 представлена зависимость элементов ковариационной матрицы вектора параметров от объема выборки. [Комментарий?].

На графике 3 представлена зависимость элементов матрицы Фишера от объема выборки. [Комментарий?].

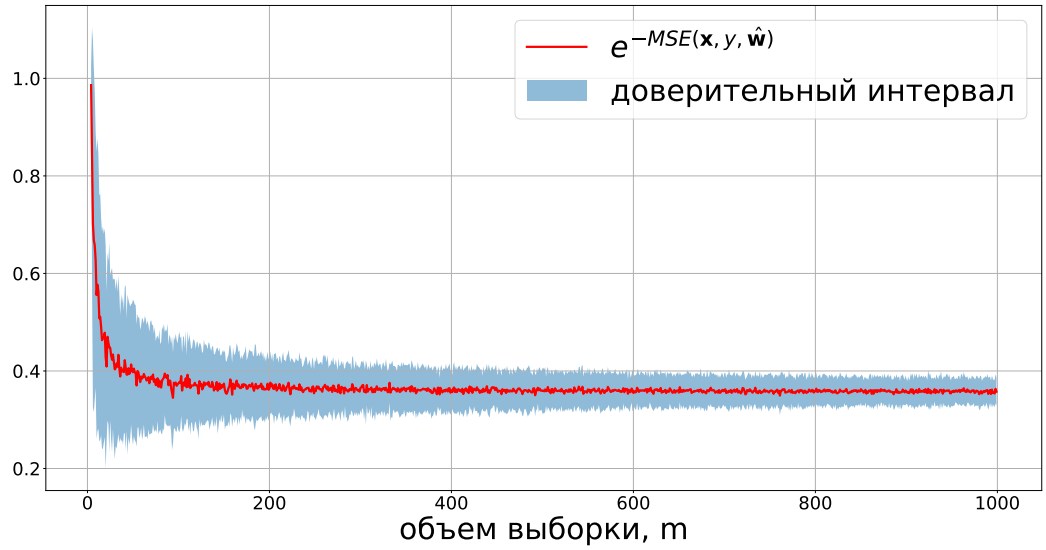



**Рис. 1** Зависимость значения функции ошибки от объема выборки

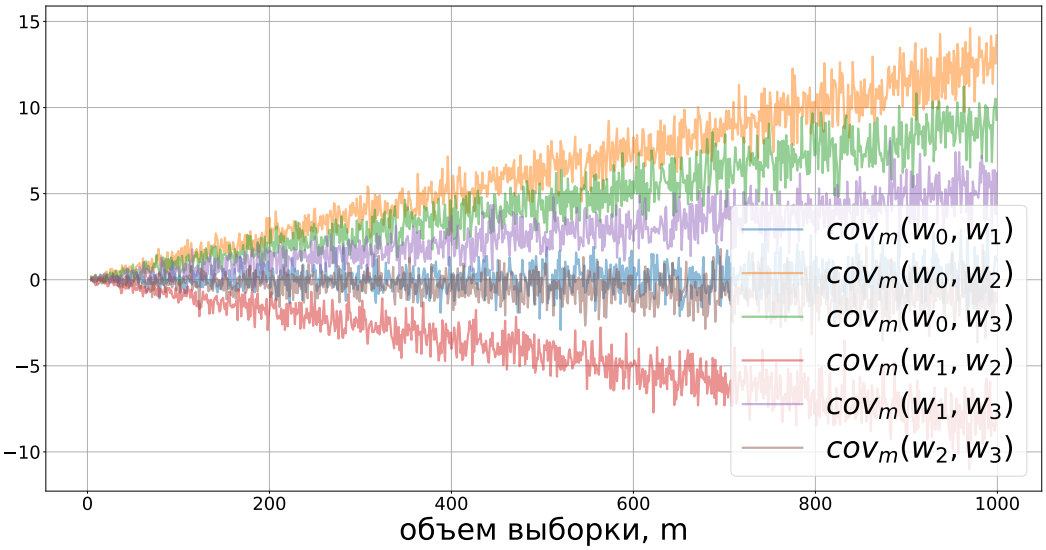



**Рис. 2** Зависимость значения элементов ковариационной матрицы от объема выборки

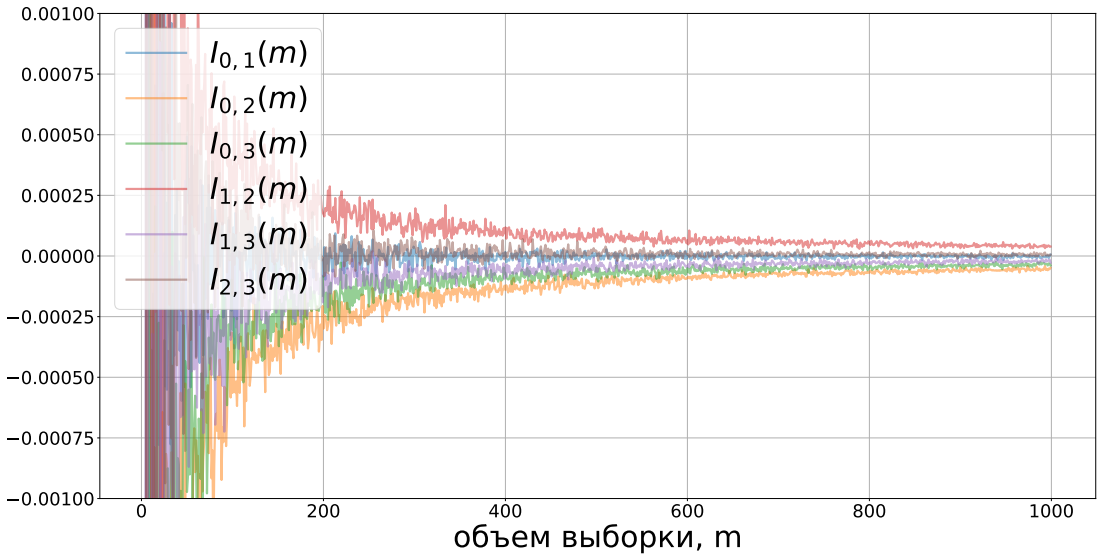



**Рис. 3** Зависимость значения элементов матрицы Фишера от объема выборки

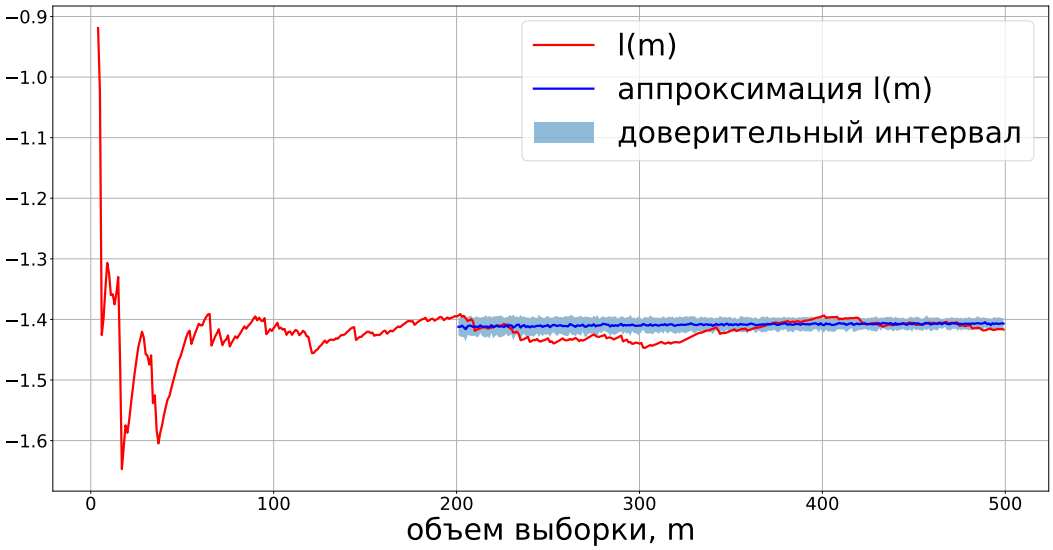



**Рис. 4** Аппроксимация функции l(m) при $m_0 = 200$

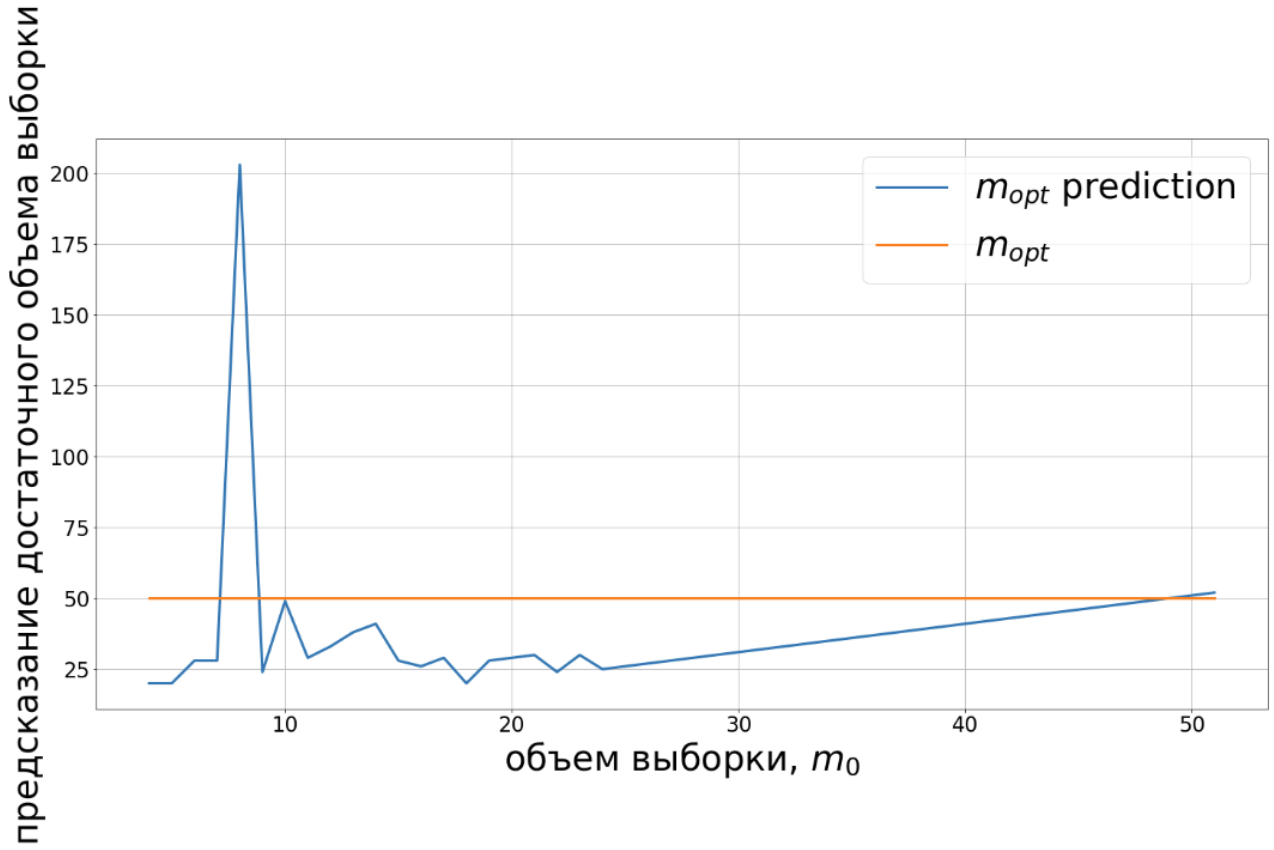



**Рис. 5** Аппроксимация оптимального размера выборки при разных $m_0$ (по волатильности аппроксимации функции l(m))

Пусть $n = 10$. Проверим, как меняется качество модели в зависимости от количества используемых признаков. Для отбора нужного количества признаков используем Лассо регрессию с функцией ошибки:

$$MSE(\mathbf{x}, y, \mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \tau\|\mathbf{w}\|_1.$$

Используется тот факт, что Лассо регрессия обнуляет малозначимые признаки, и чем больше коэффициент $\tau$, тем больше признаков обнуляется. Таким образом можно упорядочить признаки по значимости. Для того, чтобы выбрать $n'$ признаков с наилучшей функцией ошибки, достаточно взять $n'$ самых значимых признаков.

На графике 6 построена зависимость среднего значения элементов вектора параметров от величины $\tau$ при обучении Лассо регрессии на подвыборках размера $m_0 = 100$.

На графиках 7, 8 построена зависимость среднего значения функции ошибки от объёма выборки и от количества используемых признаков.

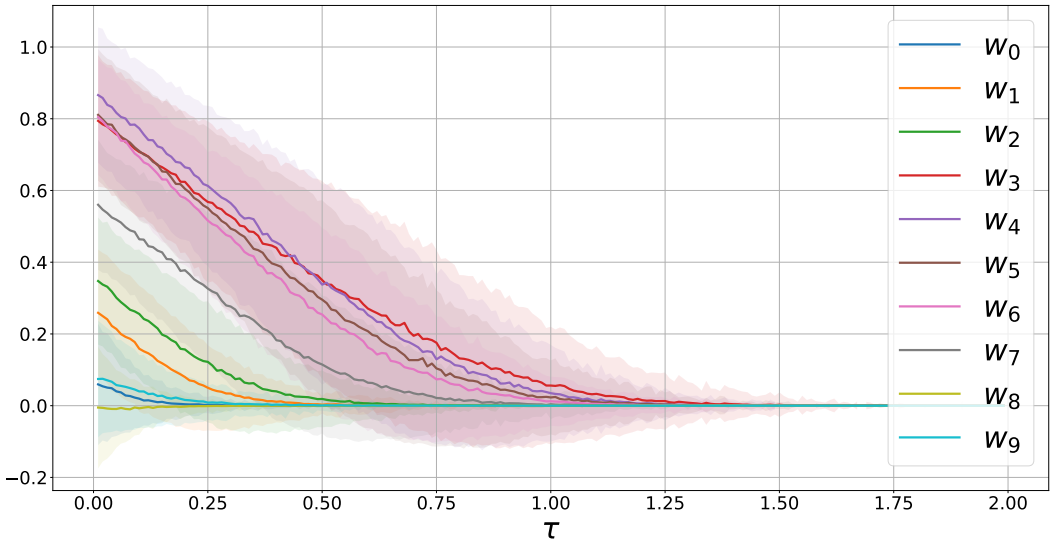



**Рис. 6** Зависимость значений элементов вектора параметров от величины $\tau$

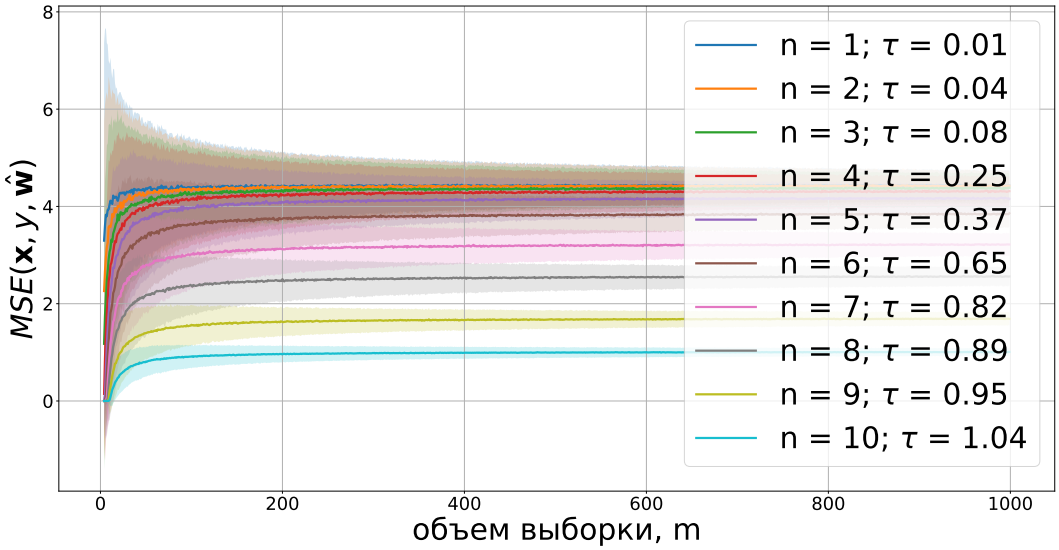



**Рис. 7** Зависимость значения функции ошибки от объёма выборки и от количества используемых признаков

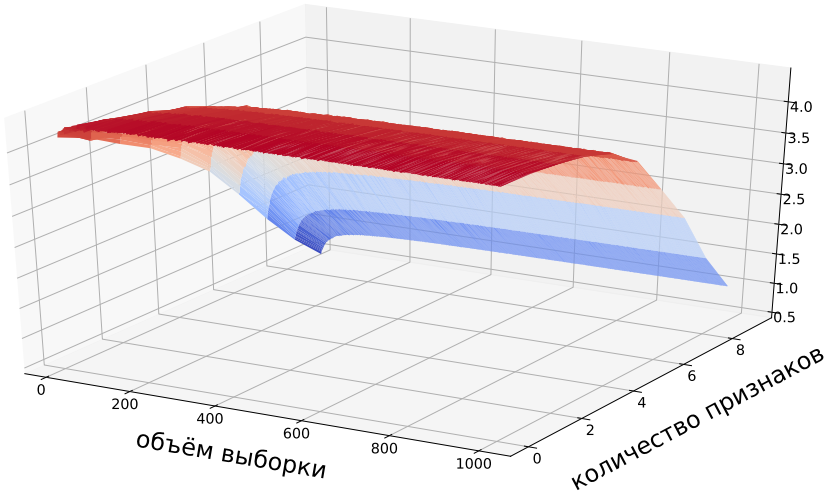



**Рис. 8** Зависимость значения функции ошибки от объёма выборки и от количества используемых признаков

## 6 Вычислительный эксперимент

Для анализа точности и эффективности предлагаемого метода был проведен вычислительный эксперимент. В качестве данных использовались выборки, описанные в таблице 1.

**Таблица 1** Описание выборок

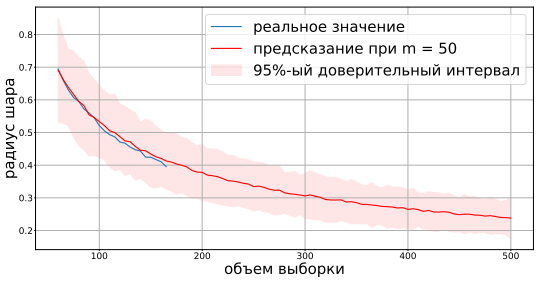
| Выборка | Тип задачи | Размер выборки | Число признаков |
|---|---|---|---|
| Servo | регрессия | 167 | 4 |
| Boston | регрессия | 506 | 13 |
| Diabetes | регрессия | 442 | 5 |

В ходе эксперимента были модифицированы критерии средней длины и среднего покрытия для линейной регрессии, а именно была построена аппроксимация функции эффективности при большем объеме выборки при помощи аппроксимации ковариационной матрицы вектора параметров через матрицу информации Фишера.

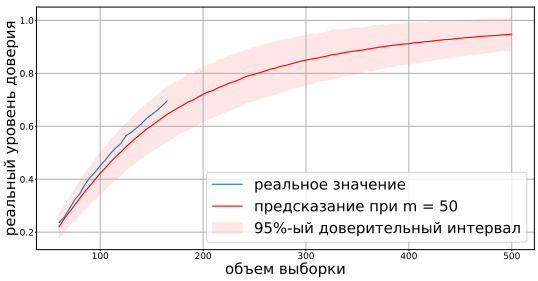
На графике 1 показана зависимость значения функции эффективности от объема выборки для разных методов при разных данных. Синим цветом обозначено посчитанное

значение при данном объеме, красным - аппроксимация при подвыборке фиксированного размера. Реальное значение функции эффективности попадает в доверительный интервал, что говорит о работоспособности предлагаемого метода.

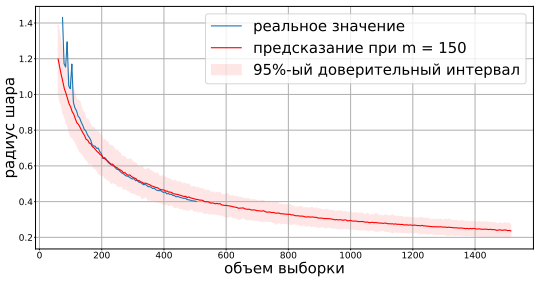
**ALC метод** **ACC метод**



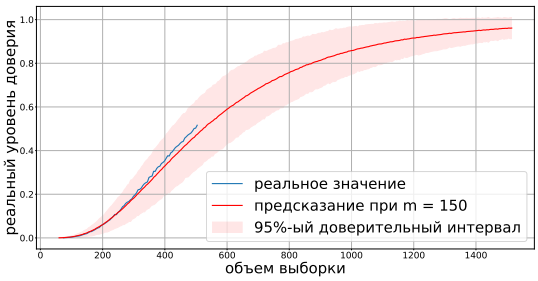

(*а*) Servo



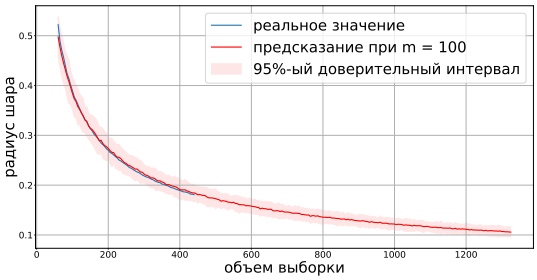

(*б*) Servo



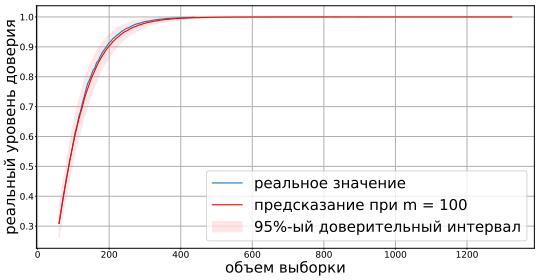

(*в*) Boston




(*г*) Boston




(*д*) Diabetes




(*е*) Diabetes

**Рис. 9** Зависимость значения функции эффективности от объема выборки

В таблицах 2, 3 представлены результаты предсказания различными методами.

**Таблица 2** Предсказание достаточного объема выборки, ALC метод

| Выборка | Реальное значение | Предсказание |
|---|---|---|
| Servo | не хватает данных | 450 |
| Boston | не хватает данных | 1370 |
| Diabetes | 235 | 240 |

**Таблица 3** Предсказание достаточного объема выборки, ACC метод

| Выборка | Реальное значение | Предсказание |
|---|---|---|
| Servo | не хватает данных | 405 |
| Boston | не хватает данных | 1410 |
| Diabetes | 235 | 245 |

## Литература

[1] *S. G. Self and R. H. Mauritsen* Power/sample size calculations for generalized linear models // Biometrics, 1988.

[2] *G. Shieh* On power and sample size calculations for likelihood ratio tests in generalized linear models // Biometrics, 2000.

[3] *G. Shieh* On power and sample size calculations for Wald tests in generalized linear models // Journal of Statistical Planning and Inference, 2005.

[4] *D. B. Rubin and H. S. Stern* Sample size determination using posterior predictive distributions // Sankhya : The Indian Journal of Statistics Special Issue on Bayesian Analysis, 1998.

[5] *Maher Qumsiyeh* Using the bootstrap for estimation the sample size in statistical experiments // Journal of modern applied statistical methods, 2002.