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, байесовские методы, бутстреп.*

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:

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

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

Выборка является простой: элементы порождены независимо из одного распределения с фиксированными неизвестными параметрами, вероятность попадания каждого элемента в выборку одинакова. Предполагается, что выборка $\mathfrak{D}$ порождена согласно следующей гипотезе: модель, порождающая данные, задается в следующем виде:

$$y_i = f(\mathbf{x}_i, \mathbf{w}, \beta), \tag{1}$$

где $\mathbf{w} \in \mathbb{W}$ — вектор параметров, β — дисперсия зависимой переменной. Зависимая переменная y аппроксимируется обобщенно линейной моделью:

$$\hat{y}_i = f(\mathbf{x}_i, \mathbf{w}) = \mu(\mathbf{w}^\top \mathbf{x}_i), \tag{2}$$

где μ — функция связи, для модели линейной регрессии:

$$\mu = id, \tag{3}$$

для логистической регрессии:

$$\mu(\mathbf{w}^\top \mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x}_i)}. \tag{4}$$

Предполагается, что при восстановлении параметров линейной регрессии (3), зависимая переменная порождается нормальным распределением:

$$p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}(f(\mathbf{x}, \mathbf{w}), \hat{\beta}),$$

где $\hat{\beta}$ — выборочная дисперсия зависимой переменной y. Для модели логистической регрессии зависимая переменная порождается бернуллиевским распределением:

$$p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{B}e\big(f(\mathbf{x}, \mathbf{w})\big).$$

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

$$L(\mathbf{w}|\mathfrak{D}) = \prod_{y,\mathbf{x}\in\mathfrak{D}} p(y|\mathbf{x}, \mathbf{w}), \quad l(\mathbf{w}|\mathfrak{D}) = \sum_{y,\mathbf{x}\in\mathfrak{D}} \log p(y|\mathbf{x}, \mathbf{w}). \tag{5}$$

Для множества индексов $\mathcal{A}$ независимой переменной $\mathbf{x} = [x_1, ..., x_n]^\top$, в векторы $\mathbf{x}_\mathcal{A}, \mathbf{w}_\mathcal{A}$ входят только те элементы, индексы которых принадлежат $\mathcal{A}$. Определим выборку $\mathfrak{D}_\mathcal{A}$ как множество независимых переменных $\mathbf{x}_\mathcal{A}$ и соответствующих им зависимых переменных y.

Для получения оптимального набора параметров и оценки вектора параметров используется принцип максимума правдоподобия:

$$\hat{\mathbf{w}}, \hat{\mathcal{A}} = \underset{\mathbf{w}\in\mathbb{W}, \mathcal{A}\subseteq\mathbb{J}}{\arg\max} L(\mathbf{w}_\mathcal{A}|\mathfrak{D}_\mathcal{A}), \tag{6}$$

где $\mathbb{J} = \{1, 2, ..., n\}$ — множество индексов.

В качестве функции ошибки используются следующие функции:

$$S_{\text{reg}}(\mathbf{x}, y, \mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2, \quad S_{\text{class}} = \sum_{i=1}^{m}(y_i \ln f(\mathbf{x}_i, \mathbf{w}) + (1 - y_i)\ln(1 - f(\mathbf{x}_i, \mathbf{w}))). \tag{7}$$

Требуется по начально заданной выборке размера $m_0 \ll m$ получить оценку минимального достаточного объёма m^*. Для введения понятия достаточности объёма выборки рассмотрим ожидаемое значение усредненного логарифма правдоподобия (5) по разным подвыборкам $\mathfrak{D}'$ размера m_0 обучающей выборки $\mathfrak{D}$ на оценке вектора параметров $\hat{\mathbf{w}}$:

$$l(m_0) = m_0^{-1} \mathop{\mathsf{E}}_{\substack{\mathfrak{D}' \subset \mathfrak{D} \\ |\mathfrak{D}'| = m_0}} l(\hat{\mathbf{w}}|\mathfrak{D}').$$

Будем считать, что объем выборки достаточен, если:

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

где δ — достаточно малое пороговое значение.

Для получения аппроксимации функции $l(m_0)$ по выборке $\mathfrak{D}_{m_0}$ введем процедуру бутстрепа:

1. Из равномерного распределения генерируется случайный вектор индексов

   $$\mathbf{i} = (i_1, ..., i_{m_0}),$$

   где:
   - $i_j \sim U(\{1, ..., m_0\})$ — эмпирический вариант,
   - $i_j \sim U(\{1, ..., m\})$ — теоретический вариант.
2. Для полученной выборки $\mathfrak{D}_{\mathbf{i}} = \{\mathbf{x}_{i_j}, y_{i_j}\}_{j=1}^{m_0}$ считается усредненный логарифм правдоподобия на оценке вектора параметров $\hat{\mathbf{w}}$:

   $$\hat{l}_k(m_0) = \frac{1}{m_0} l(\hat{\mathbf{w}}|\mathfrak{D}_{\mathbf{i}})$$

3. пп. 1-2 повторяются K раз, оценка $\hat{l}(m_0)$ функции $l(m_0)$ задается формулой:

   $$\hat{l}(m_0) = \frac{1}{K} \sum_{k=1}^{K} \hat{l}_k(m_0).$$

Для предсказания значения функции $\hat{l}(m)$ при $m > m_0$ введем семейство функций $\Phi = \{\varphi(m) = a + b \cdot m^c \mid a, b \in \mathbb{R}, c \in (-\infty, 0)\}$ Функция $\hat{l}$ приближается функцией вида:

$$\hat{\varphi} = \arg\min_{\varphi \in \Phi} \text{MAPE}(\hat{l}, \varphi, 1, m_0), \tag{8}$$

где MAPE — средняя абсолютная процентная ошибка:

$$\text{MAPE}(\psi, \varphi, m_1, m_2) = \frac{1}{m_2 - m_1 + 1} \sum_{i=m_1}^{m_2} \frac{|\varphi(i) - \psi(i)|}{\psi(i)}. \tag{9}$$

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 ε.

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 α — 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 α — 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 $\hat{\mathbf{w}}$:

$$\mathsf{E}\hat{\mathbf{w}} = \mathbf{m}, \quad \mathsf{D}\hat{\mathbf{w}} = \mathbf{I}^{-1}(\mathfrak{D}_m), \tag{10}$$

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

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

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

Оценка вектора параметров $\hat{\mathbf{w}}$ является решением задачи (6):

На рис. 1 отображена зависимость значения функции ошибки от объема выборки. Видно, что при достаточно больших m значение функции ошибки почти не меняется, и увеличение объёма выборки не приводит к улучшению модели.

Зависимость посчитана с помощью метода бутстреп:

1. Генерируются случайные непересекающиеся подвыборки $\mathfrak{D}_\mathcal{L}, \mathfrak{D}_\mathcal{T} \subset \mathfrak{D}$, $|\mathfrak{D}_\mathcal{L}| = m$, $|\mathfrak{D}_\mathcal{T}| = 50$.
2. Вычисляется оценка вектора параметров $\hat{\mathbf{w}}$ как решение оптимизационной задачи (6) на выборке $\mathfrak{D}_\mathcal{L}$.
3. Считается функция ошибки для задачи регрессии (7) на оценке вектора параметров $\hat{\mathbf{w}}$ на выборке $\mathfrak{D}_\mathcal{T}$:
4. пп. 1-3 повторяются K раз, итоговая оценка равна среднему значению среди полученных оценок функций ошибок.

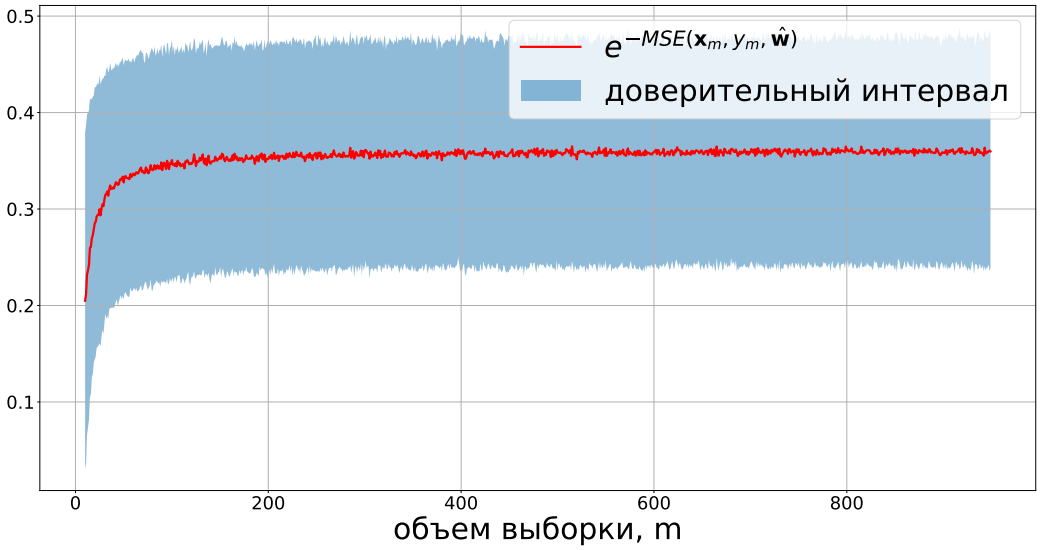


Рис. 1 Зависимость значения функции ошибки от объема выборки

5.1 Аппроксимация $\hat{l}(m, n)$

На рис. 2 представлена функция $\hat{l}(m, n)$, посчитанная в теоретическом варианте.

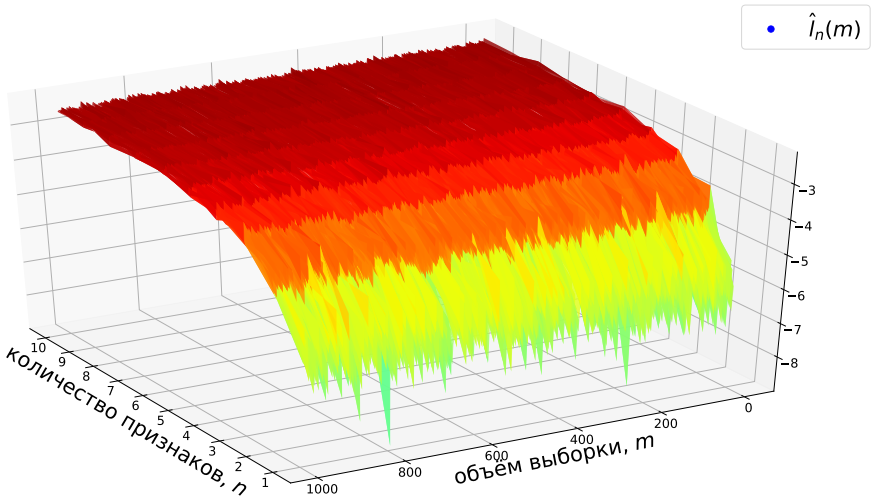


Рис. 2 Зависимость значения функции $\hat{l}$ от объема выборки m и количества параметров n.

На рис. 3 представлена аппроксимация $\varphi(p, \hat{A}(m, n), m, n) \sim \hat{l}(m, n)$ при $m_0 = 100$.

На рис. 4 представлены функции ошибки MAPE для 4 функций $\varphi \in \Phi$, аппроксимирующих $\hat{l}$, посчитанную в теоретическом и эмпирическом вариантах, с использованием или без использования ковариационной матрицы $\hat{A}$ вектора параметров $\hat{\mathbf{w}}$.

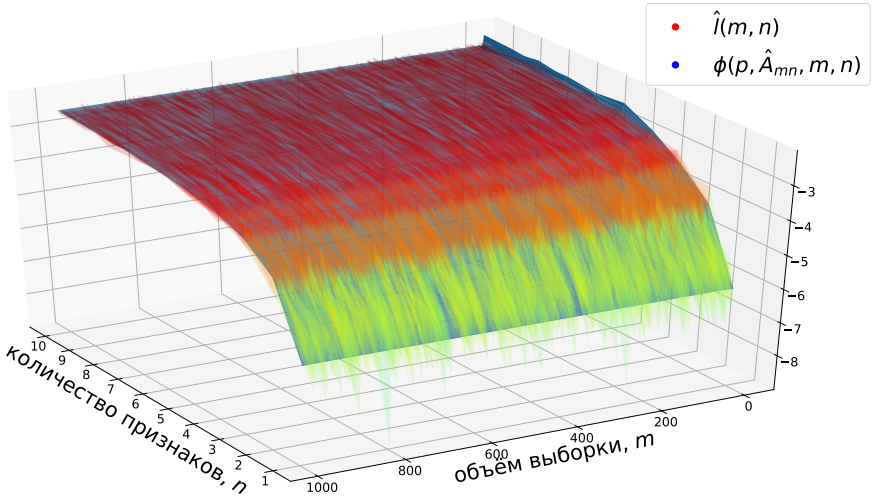


Рис. 3 Функции $\hat{l}(m,n)$ и $\varphi(p,\hat{A}(m,n),m,n)$

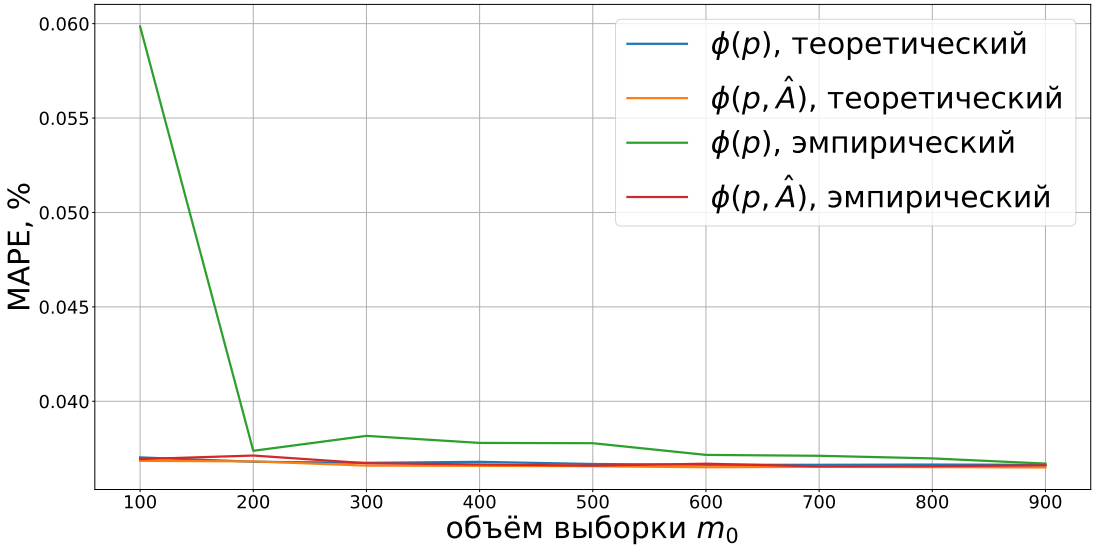


Рис. 4 Функции $\hat{l}(m,n)$ и $\varphi(p,\hat{A}(m,n),m,n)$

5.2 Использование оценки $\hat{A}$ для вычисления объёма m_0

На рис. 5 представлена зависимость элементов ковариационной матрицы вектора параметров от объема выборки. [Комментарий?]. Зависимость посчитана с помощью метода бутстреп:

1. Генерируется случайная подвыборка $\mathfrak{D}_{\mathcal{L}} \subset \mathfrak{D}$, $|\mathfrak{D}_{\mathcal{L}}| = m$
2. Вычисляется оценка вектора параметров $\hat{\mathbf{w}}_k$ как решение оптимизационной задачи (6) на выборке $\mathfrak{D}_{\mathcal{L}}$.
3. Вычисляется оценка ковариационной матрицы параметров $\hat{A}_k$ по формуле:
$$\hat{A}_k = \frac{1}{n}(\hat{\mathbf{w}}_k - \overline{\hat{\mathbf{w}}})^\top(\hat{\mathbf{w}}_k - \overline{\hat{\mathbf{w}}}). \tag{11}$$
4. пп. 1-3 повторяются K раз, итоговая оценка равна среднему значению среди полученных оценок матрицы ковариации (11).

На рис. 6 представлена зависимость элементов матрицы Фишера от объема выборки. Оценка информационной матрицы Фишера $\hat{I}$ выражена через оценку ковариационной матрицы по формуле (10).

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:

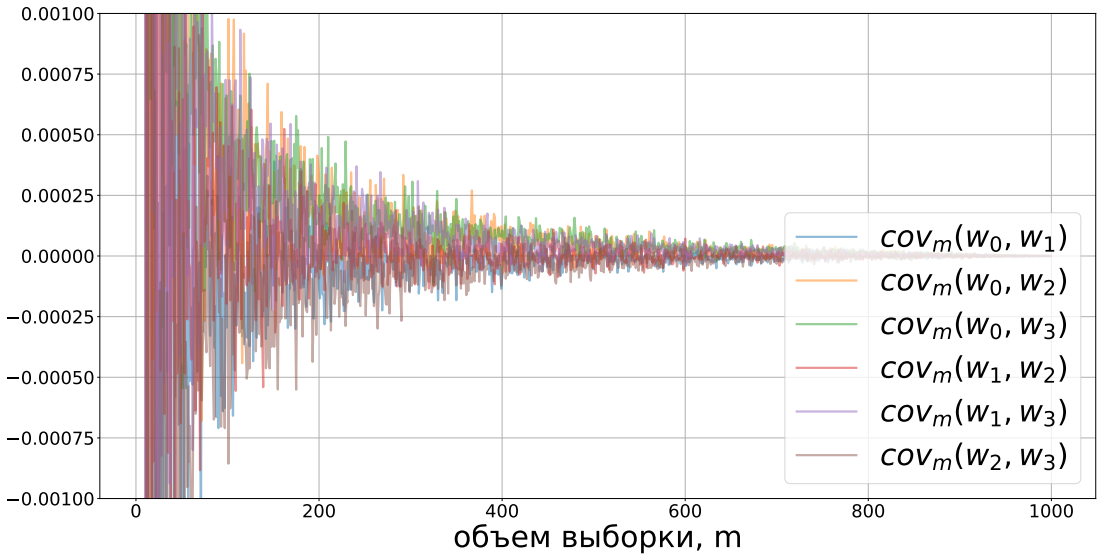


Рис. 5 Зависимость значения элементов ковариационной матрицы от объема выборки

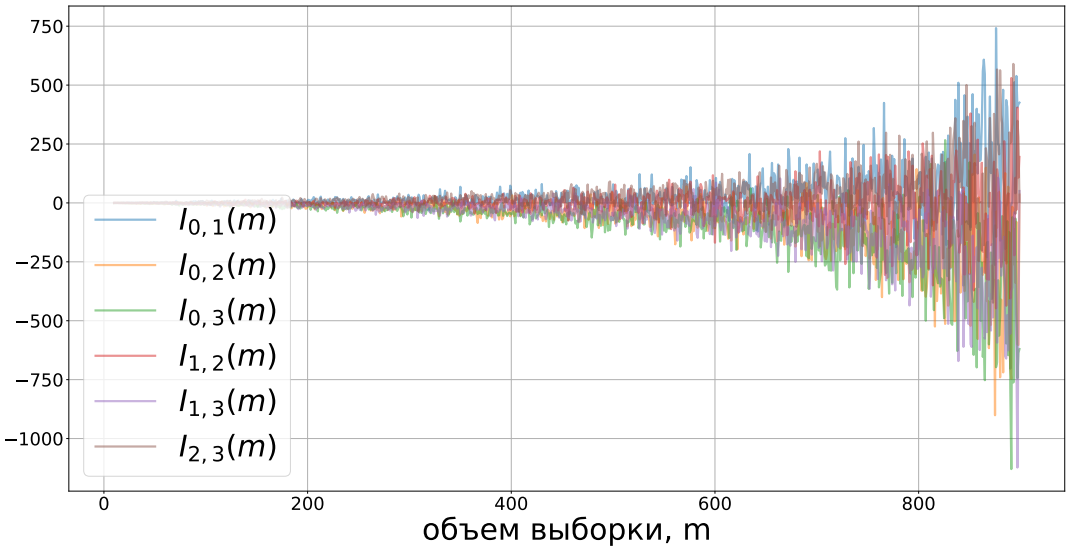


Рис. 6 Зависимость значения элементов матрицы Фишера от объема выборки

$$MSE(\mathbf{x}, y, \mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \tau\|\mathbf{w}\|_1. \tag{12}$$

Используется тот факт, что Лассо регрессия обнуляет малозначимые признаки, и чем больше коэффициент τ, тем больше признаков обнуляется. Таким образом можно упорядочить признаки по значимости. Для того, чтобы выбрать набор признаков $\mathcal{A}_{n'}$ мощности $|\mathcal{A}_{n'}| = n'$ с наилучшей функцией ошибки, достаточно взять n' самых значимых призна-

ков. Для получения оценки $\hat{\mathbf{w}}$ на наборе признаков $\mathcal{A}_{n'}$ решается оптимизационная задача (6).

На рис. 7 построена зависимость среднего значения элементов вектора параметров $\hat{\mathbf{w}}$ от величины τ при обучении Лассо регрессии (6), (12) на подвыборках размера $m_0 = 100$.

На рис. 8, 9 построены зависимости матожидания и дисперсии функции $e^{-MSE(\mathbf{x},y,\hat{\mathbf{w}})}$ от объёма выборки и от количества используемых признаков.

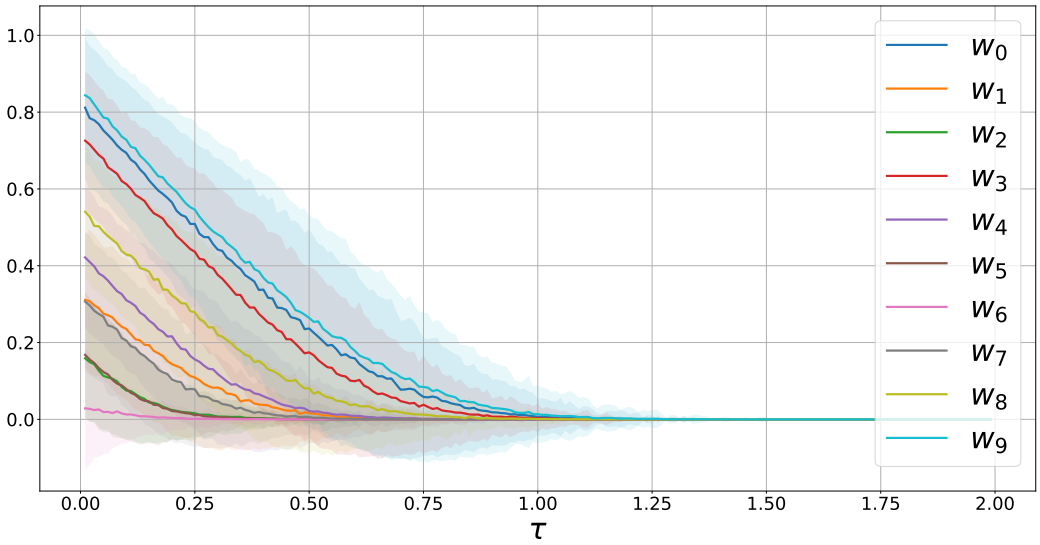


Рис. 7 Зависимость значений элементов вектора параметров от величины τ

На рис. 10 построена зависимость значения элементов оценки ковариационной матрицы параметров $\hat{A}$ от количества используемых признаков n' при обучении модели линейной регрессии (6), (**??**). Можно заметить, что ковариация между парой признаков $\mathbf{w}_i, \mathbf{w}_j$, попавших в набор используемых признаков $\mathcal{A}_{n'}$, слабо зависит от n' и остается практически неизменной.

На рис. 11 построена зависимость значения элементов оценки ковариационной матрицы параметров $\hat{A}$ от коэффициента регуляризации τ при обучении Лассо регрессии (**??**), (12). Можно заметить, что ковариация $cov(\mathbf{w}_i, \mathbf{w}_j)$ обратно пропорциональна величине τ.

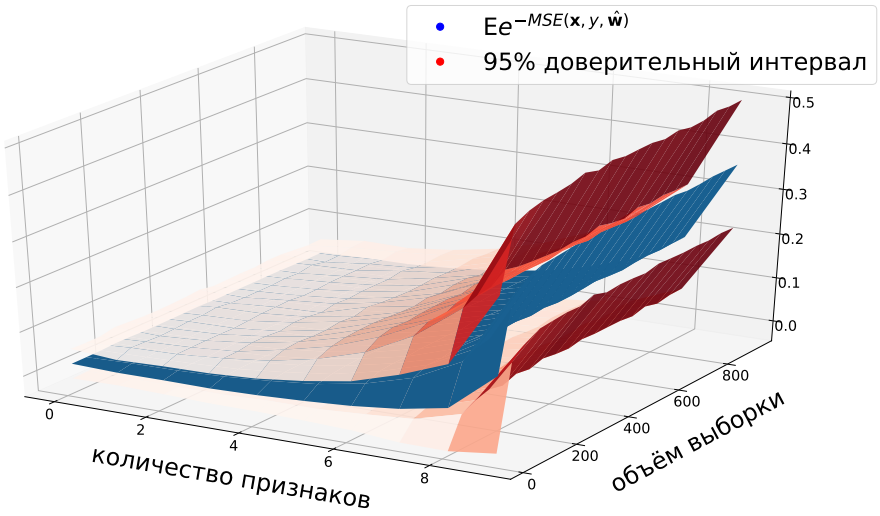


Рис. 8 Зависимость матожидания значения $e^{-MSE(\mathbf{x},y,\hat{\mathbf{w}})}$ от объёма выборки и от количества используемых признаков

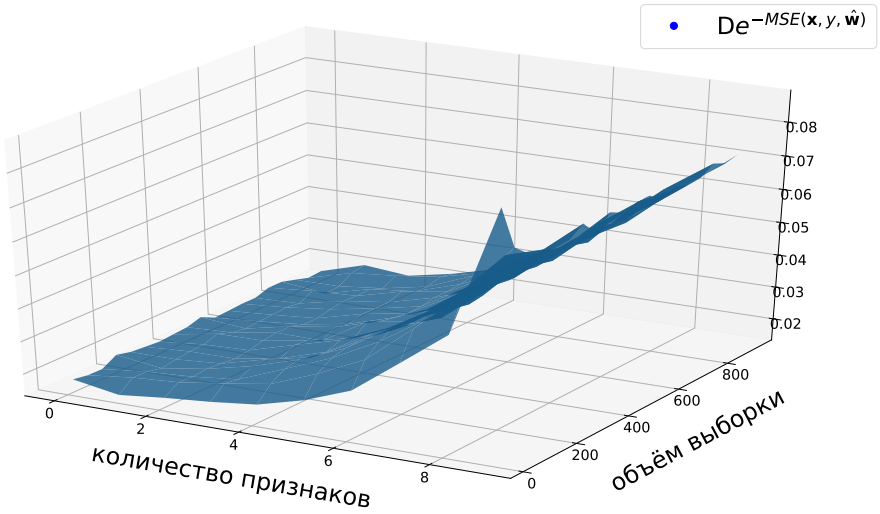


Рис. 9 Зависимость дисперсии значения $e^{-MSE(\mathbf{x},y,\hat{\mathbf{w}})}$ от объёма выборки и от количества используемых признаков

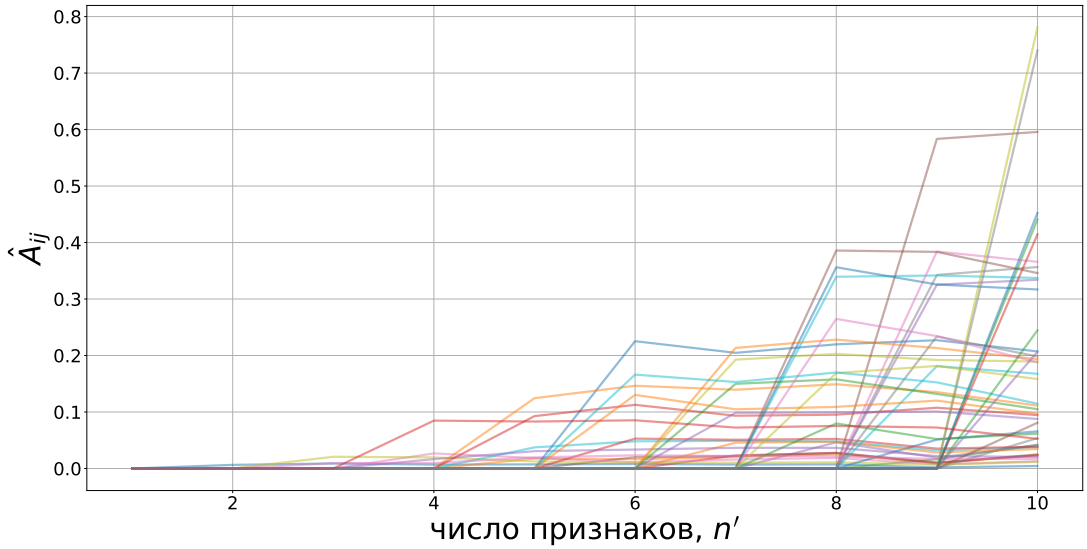


Рис. 10 Зависимость значения элементов ковариационной матрицы $\hat{A}$ от числа используемых признаков n, $m = 100$.

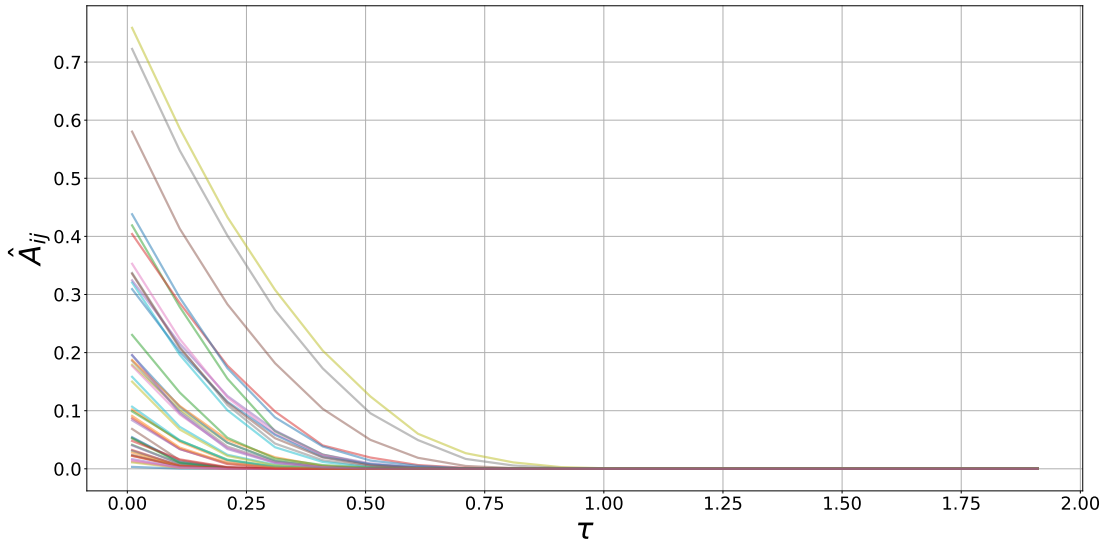


Рис. 11 Зависимость значения элементов ковариационной матрицы $\hat{A}$ от коэффициента τ, $m = 100$.

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.

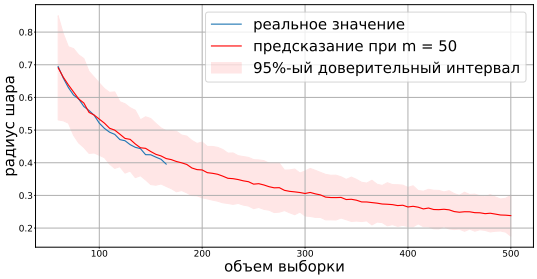
В таблицах 2, 3 представлены результаты предсказания различными методами.

Литература

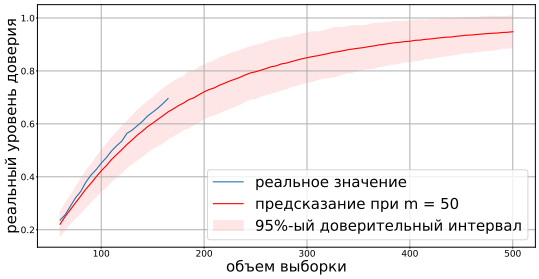
[1] *S. G. Self and R. H. Mauritsen* Power/sample size calculations for generalized linear models // Biometrics, 1988.

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

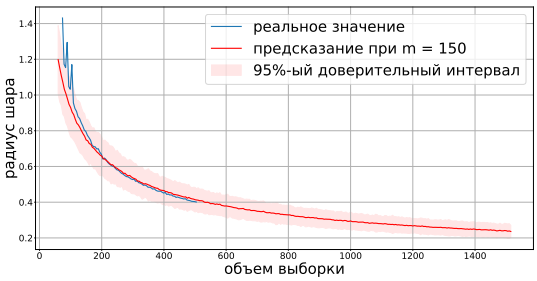
ALC метод **ACC метод**



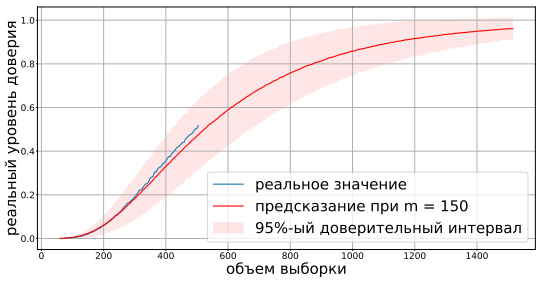
(*а*) Servo



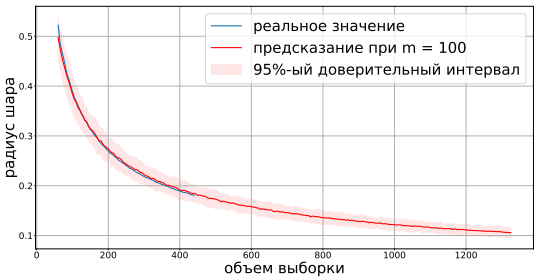
(*б*) Servo



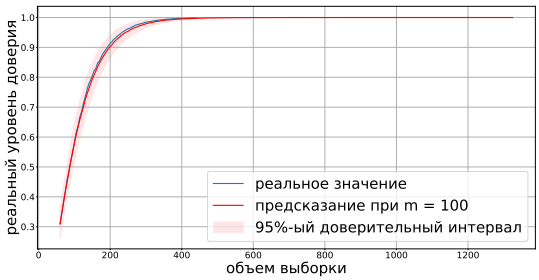
(*в*) Boston



(*г*) Boston



(*д*) Diabetes



(*е*) Diabetes

Рис. 12 Зависимость значения функции эффективности от объема выборки

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

[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.

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