**УДК 519.237.5**

ПРОЦЕДУРЫ ВЫБОРА НАИЛУЧШЕГО РЕГРЕССИОННОГО УРАВНЕНИЯ

*А. В. Быкова, О. Н. Канева*

Омский государственный технический университет, г. Омск, Россия

*В ходе проведения исследования были изучены процедуры выбора наилучшего регрессионного уравнения, проведен их анализ. Был разработан и реализован программный продукт для выбора наилучшего регрессионного уравнения.*

*Ключевые слова: регрессия, предиктор, наилучшее регрессионное уравнение, МГУА, МНК, полином Колмогорова-Габора.*

В работе рассмотрено 10 различных процедур выбора наилучшего регрессионного уравнения.

1) Метод всех возможных регрессий. Данный метод требует построения каждого из всех возможных регрессионных уравнений с переменными *Zi*. Поскольку для каждой *Zi* есть всего две возможности: либо входить, либо не входить в уравнение, то всего будет 2*i*уравнений.

2) Метод выбора «наилучшего подмножества» предикторов. В данном методе обрабатывается только часть всех возможных регрессий при определении наилучшего набора, включающего *K* уравнений, так называемого «*K*-подмножества» [1].

3) ПРЕСС – это комбинация метода всех возможных регрессий, анализа остатков и метода перепроверки.

4) Гребневая регрессия. Процедура используется, когда имеются значительные корреляции между разными предикторами, входящими в модель, и оценки параметров становятся неустойчивыми.

5) Регрессия на главных компонентах. В данном методе проблему мультиколлинеарности можно попытаться обойти используя в качестве новых переменных некоторые линейные комбинации исходных переменных, выбранные так, чтобы корреляции между ними были малы или вообще отсутствовали.

6) Регрессия на собственных значениях **–** это развитие регрессии на главных компонентах с расширенной матрицей данных, содержащей центрированные и нормированные предикторные переменные, дополненной центрированными и нормированными значениями отклика.

7) Ступенчатый регрессионный метод. После получения регрессионного уравнения для переменной *X*, наиболее сильно коррелированной с *Y*, находят остатки. Эти остатки рассматриваются как значения отклика, и строится регрессия этого отклика на предикторную переменную *X*, которая наиболее сильно коррелирована с этим новым откликом.

8) Метод исключения. Данный метод более экономичен, чем метод всех регрессий, поскольку в нем делается попытка исследовать только наилучшие регрессионные уравнения, содержащие определенное число переменных.

9) Шаговый регрессионный метод. Данный метод представляет собой попытку прийти к тем же результатом, что и метод исключения, действуя в обратном направлении, т. е. включая переменные по очереди в уравнение до тех пор, пока уравнение не станет удовлетворительным. Порядок включения определяется с помощью частного коэффициента корреляции как меры важности переменных, еще не включенных в уравнение[2].

Для программной реализации выбора наилучшей регрессионной модели было решено использовать процедуру группового учета аргументов.

10) Метод группового учета аргументов. Целью данного метода является получение модели в результате перебора моделей из индуктивно-порождаемого множества. Каждая модель настраивается – методом наименьших квадратов находятся значения параметров. Из моделей-претендентов выбираются лучшие в соответствии с выбранным критерием. [3]

Выбирается общий вид перебираемых моделей с помощью полинома Колмогорова-Габора:

$$Y \left(x\_{1}, …,x\_{n}\right)=a\_{0}+\sum\_{i=1}^{n}a\_{i}x\_{i}+\sum\_{i=1}^{n}\sum\_{j=i}^{n}a\_{ij}x\_{i}x\_{j}+\sum\_{i=1}^{n}\sum\_{j=i}^{n}\sum\_{k=j}^{n}a\_{ijk}x\_{i}x\_{j}x\_{k}+…$$

Для двух факторов количество построенных уравнений регрессии по полиному Колмогорова-Габора равно 31, для 3 факторов – 1023. Для 4 факторов количество моделей равно 131071. Так как число моделей велико, рассчитать все значения становится достаточно затруднительно, для этого была предложена реализация метод группового учета аргументов на языке программирования С#.

На вход в программу поступают массивы значений переменных из файлов формата csv. На выходе пользователь видит коэффициенты регрессии, оценки качества, а также построенный график наилучшей модели. На рисунках 1 и 2 представлены результаты работы программы для рядов с двумя и тремя факторами соответственно.



Рисунок 1 – Результат работы программы для двух факторов



Рисунок 2 – Результат работы программы для трех факторов

Для оценки качества модели используется средняя ошибка аппроксимации, которая представляет собой среднее относительное отклонение расчетных значений от наблюдаемых:

$MAPI=\frac{1}{n}\sum\_{}^{}\left|\frac{y\_{i}-\hat{y}\_{i}}{y\_{i}}\right|100\%$. (1)

Построенное уравнение регрессии можно считать удовлетворительным, если величина *MAPI* не превышает 8–10 %.

Точность построенной модели регрессии можно оценить по средней квадратической ошибке:

$MSE=\sqrt{\frac{\sum\_{}^{}(\hat{y}\_{i}-y\_{i})^{2}}{n}}$. (2)

Для анализа общего качества оцененной линейной регрессии используют обычно коэффициент детерминации $R^{2}$, называемый также квадратом коэффициента множественной корреляции. Он характеризует долю вариации (разброса) зависимой переменой, объясненной с помощью данного уравнения. Коэффициент детерминации рассчитывается по формуле:

$R^{2}=1-\frac{\sum\_{i}^{}e\_{i}^{2}}{\sum\_{i}^{}\left(y\_{i}-\overbar{y}\right)^{2}}$ , (3)

На реальных данных были построены модели, найдены коэффициенты регрессии и вычислены прогнозируемые значения уравнений. Найдены оценки качества, построенных моделей, с помощью которых можно было выявить наилучшее регрессионное уравнение.

**Библиографический список**

1. Ханк Д. Бизнес-прогнозирование [Текст] / Д. Ханк, А. Райтс, Д. Уичерн. – 7-е изд. – М., СПб., Киев : Вильямс, 2003. – 656 с.
2. Дрейпер, Н. Прикладной регрессионный анализ [Текст] : пер. с англ. Ю. П. Адлером, В. Г. Горским. / Н. Дрейпер, Г. Смит. – книга 2, 2-е изд. – М. : Финансы и статистика, 2012. – 304 с.
3. Профессиональный информационно-аналитический ресурс [Электронный ресурс]. – Режим доступа: <http://www.machinelearning.ru/wiki>, свободный. – Загл. с экрана.