УДК 004.021

Студ. П. С. Шенец
Науч. рук. доцент, к.т.н. Н. В. Пацей
(кафедра программной инженерии, БГТУ)

РЕГРЕССИОННЫЙ АНАЛИЗ ПРИ РЕШЕНИИ ЗАДАЧИ ПРОГНОЗИРОВАНИЯ

Линейная и логистическая регрессия–наиболее известные виды регрессии. Конечно, они являются наиболее важными среди всех форм регрессионного анализа, но не единственными. Каждый тип регрессии имеет свое значение и определенную область применения. Целью данной работы является анализ семи наиболее часто используемых форм регрессии для их применяя в соответствии областью и условиями использования.

Регрессионный анализ – это форма техники прогнозирующего моделирования, которая исследует взаимосвязь между зависимой (целевой) и независимой (-ыми) переменной (-ыми) (предиктором)[1]. Этот метод используется для прогнозирования, моделирования временных рядов и определения причинно-следственной связи между переменными. Например, взаимосвязь между погодными условиями и количеством ошибок в ходе передачи информации по спутниковым каналам лучше всего изучать с помощью регрессии.

Преимущества использования регрессионного анализа заключаются в следующем:

1) установление значимых отношения между зависимой и независимой переменной;
2) определение силы воздействия нескольких независимых переменных на зависимую переменную;
3) сравнение влияния переменных, измеренных в разных масштабах.

Методы регрессии, используемые для прогнозирования, в основном, определяются тремя метриками (или их комбинациями): количество независимых переменных, тип зависимых переменных и форма линии регрессии[2].

Линейная регрессия – одна из первых тем, которые изучают в прогнозирующем моделировании. Зависимая переменная при этом является непрерывной, независимая переменная (переменные) может быть непрерывной или дискретной, а характер линии регрессии – линейный [2]. Линейная регрессия устанавливает связь между зависимой переменной (Y) и одной или несколькими независимыми пере-

менными (X), используя прямую линию наилучшего соответствия, представленную уравнением:

$$Y = \alpha + \beta X,$$

где α – точка пересечения, β – наклон линии, X – независимая, Y – зависимая переменная.

Основная задача стоит в том, чтобы получить наиболее подходящую линию. Задача решается методом наименьших квадратов [2]. Это наиболее распространенный метод, используемый для подгонки линии регрессии. Он рассчитывает линию наилучшего соответствия для наблюдаемых данных, сводя к минимуму сумму квадратов вертикальных отклонений от каждой точки данных до линии. Поскольку отклонения сначала возводятся в квадрат, при добавлении не происходит компенсации между положительными и отрицательными значениями (рис.1).

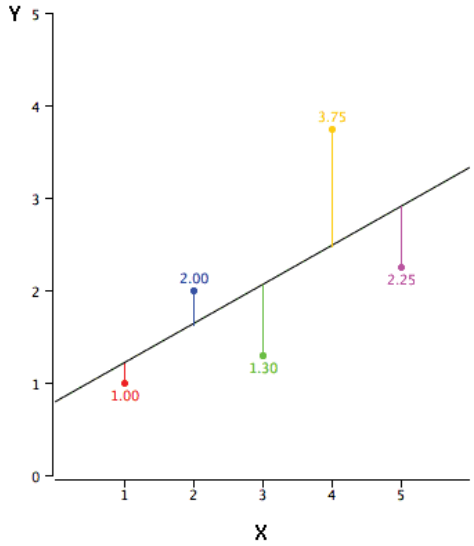

Рисунок 1 – Расчет отклонений точек данных от линии

Особенности линейной регрессии:

1) должна быть линейная связь между независимыми и зависимыми переменными;

2) возможна автокорреляция, гетероскедастичность (множественная регрессия может обладает мультиколлинеарностью), что приводит к нестабильным оценкам коэффициента;

3) чувствительна к выбросам.

Логистическая регрессия используется, чтобы найти вероятность успешного события или неудачи или когда зависимая переменная имеет двоичный характер (например: 0/1, True/False, Yes/No).

Особенности:

1) часто используется для задач классификации;

2) не требует линейных отношений между зависимыми и независимыми переменными (применяет нелинейное логарифмическое преобразование);

3) необходимо включить все значимые переменные;

4) требует больших размеров выборки;

5) независимые переменные не должны коррелировать друг с другом (т.е. не иметь мультиколлинеарность);

6) существуют разновидности – порядковая и многочленная логистическая регрессия.

Полиномиальная регрессия. Уравнение регрессии является уравнением полиномиальной регрессии, если степень независимой переменной больше 1. Уравнение, приведенное ниже, представляет полином:

$$Y = \alpha + \beta X\text{^}2.$$

В этом методе регрессии лучшая линия соответствия – кривая.

Особенности: использование полинома более высокой степени может привести к переопределению (переобучению).

Пошаговая регрессия используется при нескольких независимых переменных. В этом методе выбор независимых переменных осуществляется на основе наблюдения статистических значений: R-квадрат, t-stats и AIC метрики. Пошаговая регрессия добавляет/отбрасывает ковариации по одному на основе заданного критерия. Распространены следующие методы пошаговой регрессии: стандартная ступенчатая регрессия (добавляет и удаляет предикторы по мере необходимости для каждого шага); прямой выбор начинается с наиболее значимого предиктора в модели и добавляет переменную на каждом шаге; обратное исключение начинается со всех предикторов в модели и удаляет наименее значимую переменную для каждого шага [3].

Конечная цель этого метода – максимизировать мощность прогнозирования с минимальным количеством переменных предиктора. Это один из методов обработки набора данных большей размерности.

Регрессия хребта используется, когда данные мультиколлинеарные. Это приводит к большим значениям дисперсии, что может отклонить наблюдаемое от истинного значения. Регрессия хребта добавляет степень смещения к оценкам регрессии, что уменьшает стандартные ошибки.

Регрессия лассо способна снизить изменчивость и повысить точность моделей линейной регрессии. Однако она отличается от регрессии хребта тем, что использует абсолютные значения в корректи-

рующей функции вместо квадратов. Это приводит к тому, что некоторые оценки параметров оказываются равными нулю.

Особенности:

1) допущения аналогичны методу наименьших квадратов;

2) сокращает коэффициенты до нуля (облегчает выбор функции);

3) если группа предикторов сильно коррелирует, лассо выбирает только один из них и сокращает остальные до нуля.

Регрессия эластичных сетей – это гибрид методов регрессии лассо и хребта. Эластичная сеть используется, когда есть несколько взаимосвязанных функций.

Особенности:

1) добавляет групповой эффект в случае сильно коррелированных переменных;

2) нет ограничений на количество переменных.

Помимо семи рассмотренных, наиболее часто используемых методов регрессии, можно использовать другие модели, такие как байесовская, экологическая и робастная регрессия.

Основная проблема с регрессионными моделями заключается в том, что чем больше вариантов, тем сложнее становится выбрать правильный.

Чтобы сравнить степень соответствия моделей, анализируются различные метрики: статистическая значимость параметров, R-квадрат, скорректированный R-квадрат, AIC, BIC и др.Важным показателем точности прогнозирования является среднее квадратичное между наблюдаемыми и прогнозируемыми значениями. Установлено, что регрессионные методы лассо, хребта и эластичных сетей хорошо работают в случае высокой размерности и мультиколлинеарности среди переменных в наборе данных.

ЛИТЕРАТУРА

1. Гайдышев И. Анализ и обработка данных. — СПб.: Питер, 2001. — 750 с.

2. Анализ данных и регрессия, Мостеллер Ф., Тьюки Д. Серия: Математико-статистические методы за рубежом. Мю- 1986. – 570 с.

3. Дрейпер Н., Смит Г. Прикладной регрессионный анализ В 2-х кн. М.: Финансы и статистика,1986. — 366 с.