Лекция. Регрессионный анализ данных

Павлова

27 апреля 2023 г.

ВОПРОСЫ:

- Понятие и виды регрессионных моделей
- Этапы регрессионного анализа данных.

1 Понятие и виды регрессионных моделей

При помощи регрессионного анализа возможно решение задачи прогнозирования и классификации. Прогнозные значения вычисляются путем подстановки в уравнение регрессии параметров значений объясняющих переменных. Решение задачи классификации осуществляется таким образом: линия регрессии делит все множество объектов на два класса, и та часть множества, где значение функции больше нуля, принадлежит к одному классу, а та, где оно меньше нуля, - к другому классу. Рассмотрим основные задачи регрессионного анализа: установление формы зависимости, определение функции регрессии, оценка неизвестных значений зависимой переменной. Установление формы зависимости. Характер и форма зависимости между переменными могут образовывать следующие разновидности регрессии:

- положительная линейная регрессия (выражается в равномерном росте функции);
- положительная равноускоренно возрастающая регрессия;
- отрицательная линейная регрессия (выражается в равномерном падении функции);
- отрицательная равноускоренно убывающая регрессия;
- отрицательная равнозамедленно убывающая регрессия.

Однако описанные разновидности обычно встречаются не в чистом виде, а в сочетании друг с другом. В таком случае говорят о комбинированных формах регрессии. Определение функции регрессии. Вторая задача сводится к выяснению действия на зависимую переменную главных факторов или

причин, при неизменных прочих равных условиях, и при условии исключения воздействия на зависимую переменную случайных элементов. Функция регрессии определяется в виде математического уравнения того или иного типа. Прогнозные значения с помощью методов регрессии вычисляются путем подстановки в уравнение регрессии параметров значений объясняющих переменных. Функция регрессии определяется в виде математического уравнения того или иного типа. Регрессионный анализ очень тесно связан с корреляционным анализом Регрессионный анализ - метод изучения статистической взаимосвязи между одной зависимой количественной зависимой переменной от одной или нескольких независимых количественных переменных - Парная (простая) линейная регрессия - модель, позволяющая моделировать взаимосвязь между значениями одной входной независимой и одной выходной зависимой переменными с помощью линейной модели, например, прямой. Линейная регрессия полезна при создании не очень сложной зависимости с небольшим количеством данных Множественная регрессия - предполагает установление линейной зависимости между множеством входных независимых и одной выходной зависимой переменных Множественная линейная регрессия является линейной комбинацией входных переменных Полиномиальная регрессия - модель регрессии используется путем задания уравнения полиномов Парная (простая) линейная регрессия — это модель, позволяющая моделировать взаимосвязь между значениями одной входной независимой и одной выходной зависимой переменными с помощью линейной модели, например, прямой.

Уравнение регрессии. Это математическая формула, применяемая к независимым переменным, чтобы лучше спрогнозировать зависимую переменную, которую необходимо смоделировать

Зависимая переменная (Y) — это переменная, описывающая процесс, который мы пытаемся предсказать или понять.

Независимые переменные (X) это переменные, используемые для моделирования или прогнозирования значений зависимых переменных. В уравнении регрессии они располагаются справа от знака равенства и часто называются объяснительными переменными.

Зависимая переменная - это функция независимых переменных.

Коэффициенты регрессии — это коэффициенты, которые рассчитываются в результате выполнения регрессионного анализа. Вычисляются величины для каждой независимой переменной, которые представляют силу и тип взаимосвязи независимой переменной по отношению к зависимой. Невязки. Существует необъяснимое количество зависимых величин, представленных в уравнении регрессии как случайные ошибки.

Различают линейные и нелинейные регрессии. Предположение линейности, т.е. предполагается, что связь между рассматриваемыми переменными является линейной. Если же на диаграмме рассеивания переменных мы видим явное отсутствие линейной связи, т.е. присутствует нелинейная связь, следует использовать нелинейные методы анализа. Предположение о нормальности остатков. Оно допускает, что распределение разницы предсказанных и наблюдаемых значений является нормальным. Для

визуального определения характера распределения можно воспользоваться гистограммами остатков.

При использовании регрессионного анализа следует учитывать его основное ограничение. Оно состоит в том, что регрессионный анализ позволяет обнаружить лишь зависимости, а не связи, лежащие в основе этих зависимостей.

Регрессионный анализ дает возможность оценить степень связи между переменными путем вычисления предполагаемого значения переменной на основании нескольких известных значений.

Уравнение регрессии выглядит следующим образом: Y=a+b*X

При помощи этого уравнения переменная Y выражается через константу a и угол наклона прямой (или угловой коэффициент) b, умноженный на значение переменной X. Константу a также называют свободным членом, а угловой коэффициент - коэффициентом регрессии или B-коэффициентом. В большинстве случав (если не всегда) наблюдается определенный разброс наблюдений относительно регрессионной прямой.

Остаток - это отклонение отдельной переменной от уравнения регрессии.

Линейная регрессия характеризуется

- полезна при создании не очень сложной зависимости с небольшим количеством данных
- чувствительна к выбросам

Нелинейные регрессии делятся на два класса: регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам, и регрессии, нелинейные по оцениваемым параметрам.

Множественная линейная регрессия предполагает установление линейной зависимости между множеством входных независимых и одной выходной зависимой переменных. Модель является линейной, является линейной комбинацией входных переменных

Построение уравнения регрессии сводится к оценке ее параметров. Для оценки параметров регрессий, линейных по параметрам, используют метод наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических минимальна.

В полиномиальной регрессии степень некоторых независимых переменных превышает 1. Зависимость между переменными имеет нелинейный характер. Полиномиальная регрессия характеризуется:

- моделирует нелинейно разделенные данные . Она более гибкая и может моделировать сложные взаимосвязи.
- требуется контроль над моделированием переменных объекта (выбор степени).

- В случае неправильного выбора степени полинома модель может быть перенасыщена.

Гребневая (ридж) регрессия В случае высокой коллинеарности переменных стандартная линейная и полиномиальная регрессии неэффективны.

Коллинеарность — это отношение независимых переменных, близкое к линейному. Наличие высокой коллинеарности можно определить несколькими путями:

- коэффициент регрессии не важен, несмотря на то, что, теоретически, переменная должна иметь высокую корреляцию с Y
- при добавлении или удалении переменной из матрицы X, коэффициент регрессии сильно изменяется
- переменные матрицы X имеют высокие попарные корреляции (посмотрите корреляционную матрицу)

Регрессия по методу «лассо» В регрессии лассо, как и в гребневой, учитывается условие смещения в функцию оптимизации для того, чтобы уменьшить коллинеарность и, следовательно, дисперсию модели.

Регрессия «эластичная сеть» – комбинация методов регрессии лассо и гребневой регрессии

Этапы регрессионного анализа

- Формулировка задачи. На этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений.
- Определение зависимых и независимых (объясняющих) переменных.
- Сбор статистических данных. Данные должны быть собраны для каждой из переменных, включенных в регрессионную модель.
- Формулировка гипотезы о форме связи (простая или множественная, линейная или нелинейная).
- Определение функции регрессии (заключается в расчете численных значений параметров уравнения регрессии)
- Оценка точности регрессионного анализа
- Интерпретация полученных результатов. Полученные результаты регрессионного анализа сравниваются с предварительными гипотезами. Оценивается корректность и правдоподобие полученных результатов.
- Предсказание неизвестных значений зависимой переменной.

Оценка неизвестных значений зависимой переменной. Решение этой задачи сводится к решению задачи одного из типов:

Оценка значений зависимой переменной внутри рассматриваемого интервала исходных данных, т.е. пропущенных значений; при этом решается задача интерполяции.

Оценка будущих значений зависимой переменной, т.е. нахождение значений вне заданного интервала исходных данных; при этом решается задача экстраполяции.

Обе задачи решаются путем подстановки в уравнение регрессии найденных оценок параметров значений независимых переменных. Результат решения уравнения представляет собой оценку значения целевой (зависимой) переменной.