Методы регрессионного анализа

Как правило, при изучении сложных систем с множественными случайными данными для исследования их стохастических связей используются различные методы математической статистики, такие как: метод сопоставления двух параллельных рядов, метод аналитических группировок, корреляционный анализ, некоторые непараметрические методы.

При этом для построения математических моделей таких систем широко применяются методы регрессионного анализа. **В них в том или ином виде анализируется влияния одной или нескольких независимых переменных $\boldsymbol{X = (x_1, x_2, \dots, x_n)}$ на зависимую переменную $\boldsymbol{Y}$. Независимые переменные иначе называют регрессорами или предикторами, иногда факторными переменными, а зависимые переменные – критериальными или результативными.** При этом одна из независимых переменных может изменяться, в то время как другие независимые переменные могут оставаться фиксированными. Терминология зависимых и независимых переменных отражает лишь математическую зависимость (корреляцию) переменных, т.е. математическую постановку задачи. Анализ осуществляется статистическими методами.

Регрессия (лат. regressio - обратное движение, переход от более сложных форм развития к менее сложным, типизированным) - одно из основных понятий в теории вероятности и математической статистике, выражающее зависимость среднего значения случайной величины от значений некоторой другой величины или нескольких величин (или условное математическое ожидание). Это понятие было введено английским исследователем, географом, антропологом и психологом Фрэнсисом Гальтоном в 1886 г. Кстати, именно он является - основателем дифференциальной психологии и психометрики, статистики применительно к биологии.

Пусть имеются — случайные величины с заданным совместным распределением вероятностей.

$$(Y, x_1, x_2, \dots, x_n) \rightarrow P(Y, x_1 \epsilon B_1, x_2 \epsilon B_2, \dots, x_n \epsilon B_n),$$

При этом конечный набор случайных величин $X = (x_1, x_2, \dots, x_n)$, заданных на одном и том же вероятностном пространстве, называют

случайным вектором или многомерной случайной величиной. Это будут наши исходные данные. Совместным распределением этих случайных величин или распределением случайного вектора называют вероятности $P(Y, x_1 \epsilon B_1, x_2 \epsilon B_2, \dots, x_n \epsilon B_n)$, где множества $(B_1, B_2, \dots, B_n) \in (-\infty, +\infty)$.

Если для каждого i-ого набора значений $X_i = (x_1, x_2, \dots, x_n), i = 1, \dots, p$,определено условное математическое ожидание $M[(Y|x_1, x_2, \dots, x_n)]$, то тогда мы будем иметь **уравнение регрессии (целевую функцию)** вида:

$$f(x_1, x_2, \dots, x_n) = M[(Y|x_1, x_2, \dots, x_n)]. \quad (1)$$

График этой функции называется - **линией регрессии**

Зависимость Y от X проявляется в изменении средних значений Y при изменении наборов $X_i, i = 1, \dots, p$. При каждом фиксированном наборе значений величина Y остаётся случайной величиной с определённым распределением, т.е. это не зависимость отдельных величин y от величины x.

Чтобы модель давала нам полезную информацию, которую можно использовать при нахождении причинно-следственных связей, общих для сравниваемых случаев, необходимо иметь представление о силе соответствующих связей, то есть понимать, какие из показателей влияют на результат сильнее, а какие слабее, а также насколько велико результирующее влияние всех факторов. С этой задачей и призваны справляться регрессионные модели. В этом и состоит их сильная сторона.

Регрессионный анализ направлен не просто на изучение изменений, но на сведение воедино причины и следствия. Иначе говоря, регрессионный анализ отвечает на вопрос: «Влияет ли одна или несколько переменных (потенциальных причин) на другую переменную (результат) и, если да, то в какой степени?». Данный статистический метод исследования широко используется для прогнозирования там, где его использование имеет существенное преимущество.

Общая цель работы с регрессиями.

Вообще-то, различают *математическую модель* и *регрессионную модель*. Математическая модель предполагает участие аналитика в конструировании функции, которая описывает некоторую известную закономерность. Математическая модель является интерпретируемой -

объясняемой в рамках исследуемой закономерности. При построении математической модели сначала создаётся параметрическое семейство функций, затем с помощью измеряемых данных выполняется *идентификация модели* - нахождение её параметров.

Известная функциональная зависимость объясняющей переменной и переменной отклика - основное отличие чистого математического моделирования от регрессионного анализа. Недостаток математического моделирования состоит в том, что измеряемые данные используются для верификации, но не для построения модели, вследствие чего можно получить неадекватную модель. Также затруднительно получить модель сложного явления, в котором взаимосвязано большое число различных факторов.

Регрессионная модель объединяет широкий класс универсальных функций, которые описывают некоторую закономерность. При этом для построения модели в основном используются измеряемые данные, а не знание свойств исследуемой закономерности. Такая модель часто не интерпретируема, но более точна. Это объясняется либо большим числом моделей-претендентов, которые используются для построения оптимальной модели, либо большой сложностью модели. Нахождение параметров регрессионной модели иногда называется *обучением модели*.

При построении модели необходимо:

- определить, существуют ли статистически значимые отношения между зависимой и независимыми переменными, и как они себя проявляют. Регрессия, с одной стороны, дает возможность исследователю «ухватить» общую закономерность, а с другой оставляет пространство для возможных исключений из правила (случаев, которые в закономерность не вписываются). Т.е. необходимо по какому либо критерию разработать статистическую модель этой связи и определить форму (аналитическое выражение) влияния факторных признаков на результативный;
- предсказать значения зависимой переменной (или отклика) по значениям по крайней мере одной или нескольких, независимых (или объясняющих), переменных;
- определить вклада отдельных независимых переменных в вариацию зависимой.

Недостатки регрессионного анализа:

- модели, имеющие слишком малую сложность, могут оказаться неточными,
- модели, имеющие избыточную сложность, могут оказаться *переобученными*.

При этом следует отметить, что и регрессионная, и математическая модель, как правило, задают непрерывное отображение. Требование непрерывности обусловлено классом решаемых задач: чаще всего это описание физических, химических и других явлений, где требование непрерывности выставляется естественным образом. Иногда на отображение накладываться ограничения монотонности, гладкости, измеримости, и некоторые другие.

Примеры регрессионных моделей: линейные функции, алгебраические полиномы, ряды Чебышёва, радиальные базисные функции, нейронные сети без обратной связи, например, однослойный персептрон Розенблатта и прочее.

Персептрон - математическая или компьютерная модель восприятия информации мозгом (кибернетическая модель мозга), предложенная Фрэнком Розенблаттом в 1957 году и впервые реализованная в виде электронной машины «Марк-1» в 1960 году. Это одна из первых моделей нейросетей, а «Марк-1» — первым в мире нейрокомпьютером

Регрессионный анализ используется для прогноза, анализа временных рядов, тестирования различных гипотез и выявления скрытых взаимосвязей в данных.

Таким образом, при решении задач регрессионного анализа встают следующие вопросы.

- Как выбрать тип и структуру модели, какому именно семейству она должна принадлежать?
- Какова гипотеза порождения данных, каково распределение случайной переменной?
- Какой целевой функцией оценить качество аппроксимации?
- Каким способом отыскать параметры модели, каков должен быть алгоритм оптимизации параметров?

Регрессионным анализом называется поиск такой функции f , которая описывает эту зависимость Y от X

Задача нахождения регрессионной модели нескольких свободных переменных ставится следующим образом. Задана выборка — множество $X = (x_1, x_2, \dots, x_n)$, из $x_1, x_2, \dots, x_n \epsilon B$ значений свободных переменных и множество из $Y = y_1, y_2, \dots, y_n \epsilon R$ соответствующих им значений зависимой переменной. Эти множества обозначаются как $D = \{B, R\}$, множество исходных данных $\{x_i, y_i\}$.

Задана регрессионная модель — параметрическое семейство функций $f(\overline{\beta}, x_1, x_2, \dots, x_n)$, зависящая от вектора параметров $\overline{\beta}$ и свободных переменных X . Требуется найти наиболее вероятные параметры $\overline{\beta}$.

Причем, выборка может быть даже и не функцией, а отношением.

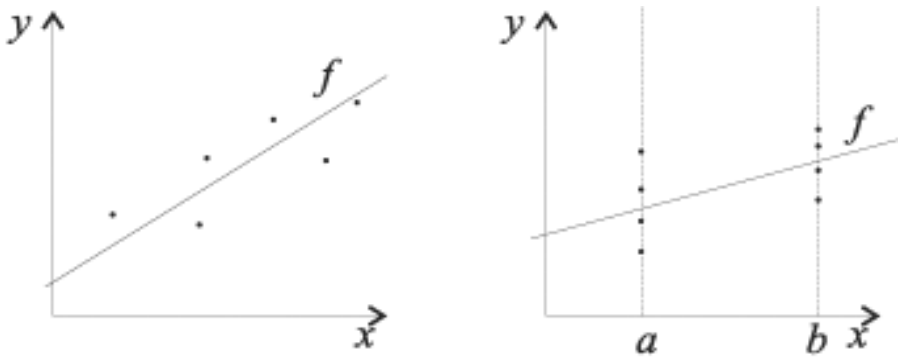


Рис.1

Это хорошо видно на примере Рис.1, где приведены примеры линий регрессии для одномерного случая. Если исходные данные, например, такие: $\{(a, 0), (a, 1), (a, 2), (a, 3), (b, 2), (b, 3), (b, 3), (b, 4)\}$, тогда в такой выборке на втором графике Рис.1 одному значению переменной x соответствует несколько значений переменной y.

Модель с одной независимой и одной зависимой переменными – это **модель парной регрессии**. Начиная со второго уровня, т.е. когда появляется x_2 и далее x_3, x_n водится понятие множественной линейной регрессии. Если объясняющих (независимых, факторных) переменных используется две или более, то говорят об использовании **модели множественной регрессии**.

Важным этапом регрессионного анализа является определение типа функции регрессии, с помощью которой характеризуется зависимость между признаками. Главным основанием должен служить содержательный анализ природы изучаемой зависимости, ее механизма. Вместе с тем теоретически обосновать форму связи каждого из факторов с результативным показателем можно далеко не всегда, поскольку исследуемые инженерно-технические или социально-экономические явления очень сложны, и факторы, формирующие их уровень, тесно переплетаются и взаимодействуют друг с другом.

Поэтому на основе теоретического анализа нередко могут быть сделаны самые общие выводы относительно направления связи, возможности его изменения в исследуемой совокупности, возможного наличия экстремальных значений и т.п. Необходимым дополнением такого рода предположений должен быть анализ конкретных фактических данных.

Приблизительно представление о функции связи можно получить на основе **эмпирической линии регрессии**. Эмпирическая линия регрессии обычно является ломанной линией, имеет более или менее значительный излом. Объясняется это тем, что влияние прочих неучтенных факторов, оказывающих воздействие на вариацию результативного признака, в средних показателях погашается неполностью, в силу недостаточно большого количества наблюдений, поэтому эмпирической линией связи для выбора и обоснования типа теоретической кривой можно воспользоваться при условии, что число наблюдений будет достаточно велико.

Одним из элементов конкретных исследований является сопоставление различных уравнений зависимости, основанное на использовании критериев качества аппроксимации эмпирических данных конкурирующими вариантами моделей. Наиболее часто для характеристики связей инженерно-технических и экономических показателей используют следующие типы функций:

1. линейная,
2. гиперболическая,
3. показательная,
4. параболическая,
5. степенная,
6. логарифмическая,
7. логистическая.

Линейный регрессионный анализ.

Термин «**линейный регрессионный анализ**» используют, когда рассматриваемая функция линейно зависит от оцениваемых параметров (от независимых переменных зависимость может быть произвольной). Теория оценивания неизвестных параметров хорошо развита именно в случае линейного регрессионного анализа. Если же линейности нет и нельзя перейти к линейной задаче, то, как правило, хороших свойств от оценок ожидать не приходится.

Линейная регрессия представляет собой линейную функцию между условным математическим ожиданием $M(Y|X = x_i)$ зависимой переменной Y и одной независимой объясняющей переменной X:

$$Y = M(Y|X = x_i) = \beta_0 + \beta_1 x_i, \qquad (2)$$

Где x_i - значения независимой переменной в i-ом наблюдении, i=1,2,…,n. Здесь принципиальной является линейность уравнения по параметрам x_i, β_1. Так как каждое индивидуальное значение y_i отклоняется от соответствующего условного математического ожидания $M(Y|X = x_i)$, тогда в данную формулу необходимо ввести случайное слагаемое ε , и, как следствие, получим:

$$Y = M(Y|X = x_i) = \beta_0 + \beta_1 x_i + \varepsilon_i. \qquad (3)$$

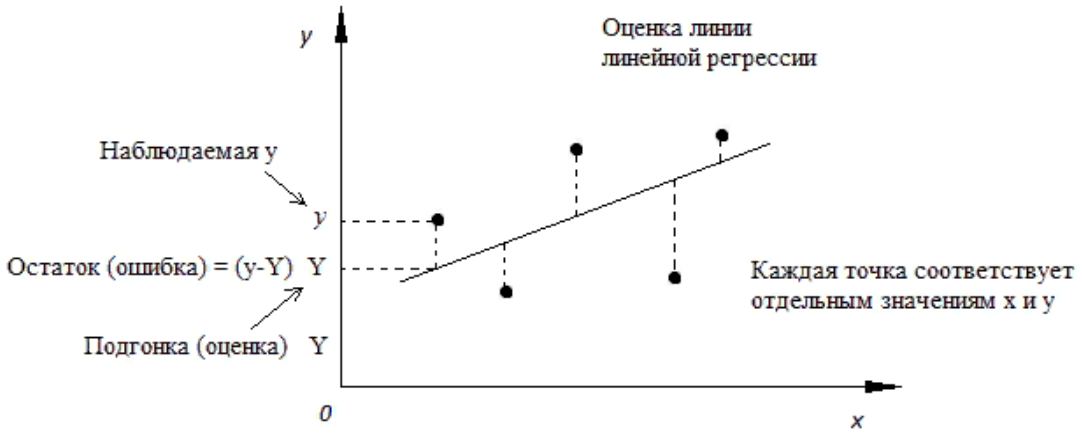


Рис.2

Или в общем случае многомерной регрессии:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \overline{\varepsilon}. \qquad (4)$$

Данное соотношение называется теоретической линейной регрессионной моделью, а $\beta_0, \dots, \beta_n$ - теоретическими параметрами (теоретическими коэффициентами) регрессии, $\overline{\varepsilon}$ - ошибка, возникающая вследствие несовпадения предсказанных и реальных значений.

Таким образом, регрессия в общем виде может быть представлена в виде суммы неслучайной и случайной составляющих.

$$Y = f(X) + \varepsilon, \qquad (5)$$

где $f = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$ — функция регрессионной зависимости, а ε — аддитивная случайная величина.

При этом, как можно видеть, в уравнениях регрессии (2)-(4) параметр β_0 показывает усредненное влияние на результативный признак неучтенных в уравнении факторных признаков (предикторов). Коэффициент регрессии β_i показывает, на сколько в среднем изменяется значение результативного признака при увеличении i-ого факторного признака на единицу собственного измерения.

Случайное отклонение ε вычисленной величины $f(X)$ от ее фактического значения Y называют невязкой.

Как мы видим, модели регрессионного анализа включают следующие переменные:

- Неизвестные параметры, обозначенные здесь как β_i, которые могут представлять собой скаляр или вектор.
- Независимые переменные, X.
- Зависимые переменные, Y.

Как уже было сказано, в различных областях науки, где осуществляется применение регрессионного анализа, используются различные термины вместо зависимых и независимых переменных, но во всех случаях регрессионная модель относит Y к функции X и β.

Для определения значений теоретических коэффициентов регрессии необходимо знать и использовать все значения переменных X и Y генеральной совокупности, что невозможно.

Задачи регрессионного линейного анализа состоят в том, чтобы по имеющимся статистическим данным $\{x_i, y_i\}, i = 1, \dots, n$ для переменных X и Y:

1.получить наилучшие оценки неизвестных параметров β_0 и β_1;

2.проверить статистические гипотезы о параметрах модели;

3.проверить, достаточно ли хорошо модель согласуется со статистическими данными.

Поэтому в целом, алгоритм регрессии является итерационным и будет продолжать перемещать линию через каждую итерацию, пытаясь найти наиболее подходящую линию, другими словами, линию с минимальной ошибкой.

Расчет мощности и объема выборки

Здесь, как правило, нет согласованных методов, касающихся числа наблюдений по сравнению с числом независимых переменных в модели. Одно из правил можно сформулировать в следующем виде:

$$N = t^k,$$

где $N -$ является размером выборки, k - число независимых переменных, а t - есть число наблюдений, необходимых для достижения желаемой точности, при условии, если модель имела только одну независимую переменную.

Например, исследователь строит модель линейной регрессии с использованием набора данных, который содержит, например, 1000 пациентов или элементов РЭС (N). Если исследователь решает, что необходимо 5 наблюдений (t), чтобы точно определить прямую (или имеется возможность провести 5 наблюдений), то максимальное число независимых переменных, которые модель может поддерживать (k), равно 4:

$$5^4 = 625 < 1000, 5^5 = 3125.$$

Предположим теперь, что вектор неизвестных параметров $\overline{\beta}$ имеет длину k. Для выполнения регрессионного анализа пользователь должен предоставить информацию о зависимой переменной Y:

1. Если наблюдаются точки N данных вида (Y, X), где $N < k$, большинство классических подходов к регрессионному анализу не могут быть выполнены, так как система уравнений, определяющих модель регрессии в качестве недоопределенной, не имеет достаточного количества данных, чтобы восстановить все $\overline{\beta}$.
2. Если наблюдаются ровно $N = k$, а функция регрессии f является линейной, то уравнение $Y = f(X, \overline{\beta})$ можно решить точно, а не приблизительно. Это сводится к решению набора N -уравнений с k -

неизвестными (элементы $\overline{\beta}$), который имеет единственное решение до тех пор, пока X линейно независим. Если f является нелинейным, решение может не существовать, или может существовать много решений.

3. Наиболее распространенной является ситуация, где наблюдается $N > k$. В этом случае имеется достаточно информации в данных, чтобы оценить уникальное значение для $\overline{\beta}$, которое наилучшим образом соответствует данным, и модель регрессии, когда применение к данным можно рассматривать как переопределенную систему в $\overline{\beta}$.

Таким образом, если, например, мы рассмотрим модель регрессии, которая имеет три неизвестных параметра: β_0, β_1 и β_2, далее предположим, что экспериментатор выполняет 10 измерений в одном и том же значении независимой переменной вектора X. В этом случае регрессионный анализ не дает уникальный набор значений. Лучшее, что можно сделать, оценить среднее значение и стандартное отклонение зависимой переменной Y. Аналогичным образом, измеряя два различных значения X, можно получить достаточно данных для регрессии с двумя неизвестными, но не для трех и более неизвестных.

Когда число измерений N больше, чем число неизвестных параметров k и погрешности измерений ε_i, то, как правило, избыток информации, содержащейся в измерениях, затем используется для статистических прогнозов относительно неизвестных параметров. Этот **избыток информации называется степенью свободы регрессии**.

Предположения классической линейной регрессии.

В классической модели линейной регрессии помимо функциональных соотношений накладываются дополнительные и весьма жесткие предположения о стохастической структуре модели.

1. Итак, для каждой наблюдаемой величины x остаток равен разнице y и соответствующего предсказанного Y. Каждый остаток может быть положительным или отрицательным. При этом $M(\varepsilon_i) = 0$.

2. Все ошибки линейно независимы друг от друга, то есть не представляется возможным выразить любой предсказатель в виде линейной комбинации остальных.

3. Остатки имеют одну и ту же вариабельность (постоянную дисперсию) для всех предсказанных величин Y, т.е. $M(\varepsilon_i^2) = \sigma^2$. Такое явление называют **гомоскедастичностью.**

4. Ошибки являются некоррелированными $M(\varepsilon_i \varepsilon_j) = 0$, то есть ковариационная матрица ошибок диагональна, и каждый ненулевой элемент на диагонали является дисперсией ошибки.

5. Часто бывает полезным предположение о явной форме ошибок, такое что $\varepsilon_i \approx N(0, \sigma^2)$ - ошибка представляет собой нормально распределенную случайную величину.

Предположение о характере распределения этой величины называется **гипотезой порождения данных.**

Некоторые из коэффициентов могут оказаться пренебрежимо малыми – незначимыми. Чтобы установить, значим коэффициент или нет, необходимо прежде всего вычислить оценку S_i^2 дисперсии, с которой он находится, и оценить итоговую погрешность:

$$\sigma_y^2 = \frac{1}{N} \sum_{i=1}^{N} S_i^2. \quad (6)$$

Одним из показателей устойчивости выбранной модели выступает стандартная ошибка коэффициентов регрессии, которая демонстрирует дисперсию независимой переменной. Это так называемые коэффициенты Стьюдента, для которых имеются табличные значения по заданному уровню значимости p:

$$t_p = \frac{|\beta_i|}{\sqrt{\sigma_i^2}}. \quad (7)$$

Для проверки воспроизводимости опытов в целом находится отношение наибольшей из оценок дисперсий к сумме всех оценок дисперсий (расчетное значение также табулированного критерий Кохрена):

$$G_p = \frac{max\ \sigma_i^2}{\sum_{i=1}^{n} \sigma_i^2}. \quad (8)$$

Если допущения линейности, нормальности и/или постоянной дисперсии сомнительны, мы можем монотонно преобразовать x или y и рассчитать новую линию регрессии, для которой эти допущения удовлетворяются (например, использовать логарифмическое преобразование или др.).

Метод наименьших квадратов

На практике все это происходит следующим образом. Предположим, что в результате опыта мы получили ряд экспериментальных точек и построили интерполяционный график зависимости y от x (рис.2), выражаемый аналитически полиномом степени $(n-1)$, так, чтобы функция в точности прошла через каждую из точек (рис.3).

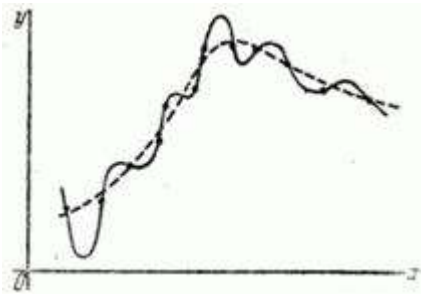


Рис.3

Обычно экспериментальные точки на таком графике располагаются не совсем правильным образом - дают некоторый «разброс», т. е. обнаруживают случайные отклонения от видимой общей закономерности. Именно подобные отклонения ε_i связаны с неизбежными при всяком опыте ошибками измерения и являются нерегулярными или случайными.

Если заранее известно, что результаты y_i измерений содержат погрешности ε_i, то естественно рассматривать **не задачу интерполяции, а использовать аппроксимацию как задачу сглаживания**, т.е. задачу построения гладкой (непрерывно дифференцируемой, дважды непрерывно дифференцируемой) функции, которая в общем случае проходила бы не через заданные точки $\{x_i, y_i\}$, $i=0,1,\ldots, n$, а вблизи них, устраняя возможные ошибки, вызванные погрешностью измерений. Иначе при каждом новом опыте мы получим новый набор данных и у нас возникнет необходимость строить новую интерполяционную кривую, которой в результате мы не сможем воспользоваться.

Если погрешность данных мала, то, конечно, используют интерполяцию, т.е. пренебрегая погрешностями единожды рассчитывают сглаживающую кривую, проходящую через каждую экспериментальную точку.

При невозможности пренебречь ошибками возникает весьма типичная для практики задача сглаживания экспериментальной зависимости, т.е. появляется необходимость в использовании методов регрессионного анализа. И вышеприведенная гипотеза порождения данных оказывается вполне справедливой.

Существует много способов определить коэффициенты регрессионной модели β_i .

Но при подобных предположениях основным и наиболее качественным способом подбора параметров β_i модели **для оценки качества модели используется критерий минимума суммы квадратов регрессионных остатков.** Он называется методом наименьших квадратов:

$$\min \sum_{i=1}^{n} \left(y_i - f\left(\vec{\beta}, x_i\right) \right)^2. \qquad (9)$$

Метод наименьших квадратов получил широкое распространение при статистической обработке экспериментальных данных, содержащих случайные ошибки.

Это, - во-первых, а, во-вторых, - часто в задачах, например, радиотехники, необходимо провести восстановление функции $f(x)$ и вычисление производной функции по зашумленным данным.

Использование интерполяции в такой ситуации нецелесообразно, т.к. интерполирующая кривая будет существенно зависеть от погрешности, а ее производная будет сильно отличаться от реальной $df(x)/dx$.

Так вот, - способ (метод) наименьших квадратов представляет собой один из наилучших методов аппроксимации. Этот метод является также и самой ранней формой регрессии. Он был опубликован Лежандром в 1805 году и Гауссом в 1809. Лежандр и Гаусс применили метод к задаче определения из астрономических наблюдений орбиты тел вокруг Солнца (в основном кометы, но позже и вновь открытые малые планеты).

Метод наименьших квадратов применяется для приближенной замены заданной функции другими более простыми функциями. Он позволяет получать наилучшую функциональную зависимость по

набору имеющихся точек (наилучшую означает, что сумма квадратов отклонений минимальна).

Пусть задан набор экспериментальных точек с координатами $(x_k, y_k), k = 0,1, \dots, K$. Согласно методу наименьших квадратов аппроксимирующую функцию выбирают таким образом, чтобы была минимальна сумма

$$S = \sum_{k=1}^{K} w_k [Y(x_k) - y_k]^2, \qquad Y(x_k) = f(\beta, x_k) \tag{10}$$

Величину w_k называют весом результата измерения. Веса обратно пропорциональны дисперсиям ошибок. В случае равноточных измерений полагают $w_1 = w_2 = \cdots = w_k = 1$.
Величина $|Y(x_k) - y_k|$ - это абсолютное значение отклонения экспериментальной точки от той, которая вычислена из функциональной зависимости $Y(X)$. Чаще всего функцию $Y(X)$ выбирают в виде алгебраического полинома:

$$Y(X) = P_v(X) = C_0 + C_1 X + C_2 X^2 + \cdots + C_v X^v. \tag{11}$$

Как мы видим, одномерная регрессия — частный случай полиномиальной регрессии.

Полиномиальная регрессия - это форма линейной регрессии, в которой взаимосвязь между независимой переменной x и зависимой переменной y моделируется как полином v -й степени. Полиномиальная регрессия может применяться в математической статистике при моделировании трендовых составляющих временных рядов.

Временной ряд — ведь это, по сути, ряд чисел, которые зависят от времени. Например, средние значения температуры воздуха по дням за прошедший год, или доход предприятия по месяцам. Цель построения модели полиномиальной регрессии в области временных рядов всё та же – прогнозирование.

В данной модели коэффициенты полинома $C_0, C_1, C_2, \dots, C_v$ находятся из критерия наименьших квадратов как

$$S = \sum_{k=1}^{K} w_k [C_0 + C_1 x_k + C_2 x_k^2 + \cdots + C_v x_k^v - y_k]^2 = min \tag{12}$$

Из курса математического анализа известно, чтобы найти минимум функции, необходимо вычислить частные производные по каждому из параметров и приравнять их к нулю. Вычисляя соответствующие производные $\partial S/\partial C_0, \partial S/\partial C_1, \dots, \partial S/\partial C_v$ и

приравнивая их к нулю, получаем так называемые нормальные уравнения:

$$\begin{aligned}
&\left(\sum_k w_k\right) C_0 + \left(\sum_k w_k x_k\right) C_1 + \ldots + \left(\sum_k w_k x_k^{\nu}\right) C_{\nu} = \sum_k w_k y_k \\
&\left(\sum_k w_k x_k\right) C_0 + \left(\sum_k w_k x_k^{2}\right) C_1 + \ldots + \left(\sum_k w_k x_k^{\nu+1}\right) C_{\nu} = \sum_k w_k x_k y_k \\
&\ldots\ldots\ldots\ldots \\
&\left(\sum_k w_k x_k^{\nu}\right) C_0 + \left(\sum_k w_k x_k^{\nu+1}\right) C_1 + \ldots + \left(\sum_k w_k x_k^{2\nu}\right) C_{\nu} = \sum_k w_k x_k^{\nu} y_k
\end{aligned} \qquad (13)$$

Это система линейных алгебраических уравнений относительно неизвестных $C_0, C_1, C_2, \ldots, C_v$. Для полинома степени v получается система из $N = v + 1$ уравнений. Количество различных сумм, которые надо вычислять для определения матрицы коэффициентов при неизвестных равно $(2v + 1)$, для определения вектора-столбца свободных членов - $(v + 1)$.

Решение такой системы нормальных уравнений может проводиться методом исключения Гаусса.

Часто заранее оказывается неизвестно, какого порядка нужно взять полином, чтобы хорошо описать экспериментальные данные.

Критерий, позволяющий считать, что полином v -ого порядка хорошо описывает экспериментальные данные, а полином $(v - 1)$ -ого порядка в этом отношении еще неудовлетворителен, зависит от вида решаемой задачи. Чаще всего в качестве такого критерия берется среднеквадратичное отклонение или максимальное абсолютное отклонение.

Метод наименьших квадратов может использоваться не только с алгебраическими полиномами, но и с функциями другого типа.

Нормальные уравнения не всегда оказываются линейными. Выбор типа функциональной зависимости, который лучше всего подходит к данным экспериментальным точкам, возлагается на исследователя.

Рекомендуется ограничиваться степенями полиномов не выше 10-ой. Использование полиномов очень высоких степеней проблематично из-за вычислительных ошибок округления.

Обобщенный метод наименьших квадратов

Существует, так называемый, обобщённый метод наименьших квадратов — метод оценки параметров регрессионных моделей, являющийся обобщением классического метода наименьших квадратов. Это происходит, когда нарушаются предположения о классической линейной регрессии. Обобщённый метод наименьших квадратов сводится к минимизации «обобщённой суммы квадратов» остатков регрессии.

Как обнаружить, что перечисленные выше условия линейной регрессии не соблюдены? Ну, во-первых, довольно часто это видно невооруженным глазом на графике.

Рассмотрим, к примеру, поведение самих остатков, например, из рис.4 следует, что они постоянны и неслучайны.

В этих случаях необходимо либо применять другую функцию, либо вводить дополнительную информацию и заново строить уравнение регрессии до тех пор, пока остатки ε_jне будут случайными величинами.

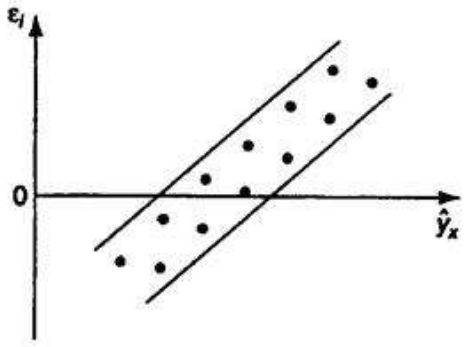


Рис. 4. Зависимость случайных остатков ε_j от теоретических значений y_x для всех x.

Для анализа следующей предпосылки МНК относительно нулевой средней величины остатков необходимо наряду с изложенным графиком рис.3 зависимости остатков ε_j от теоретических значений результативного признака y_x построить график зависимости случайных остатков ε_j и от входных факторов, включенных в регрессию x_j

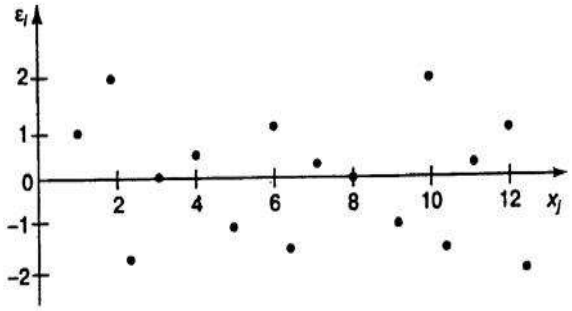
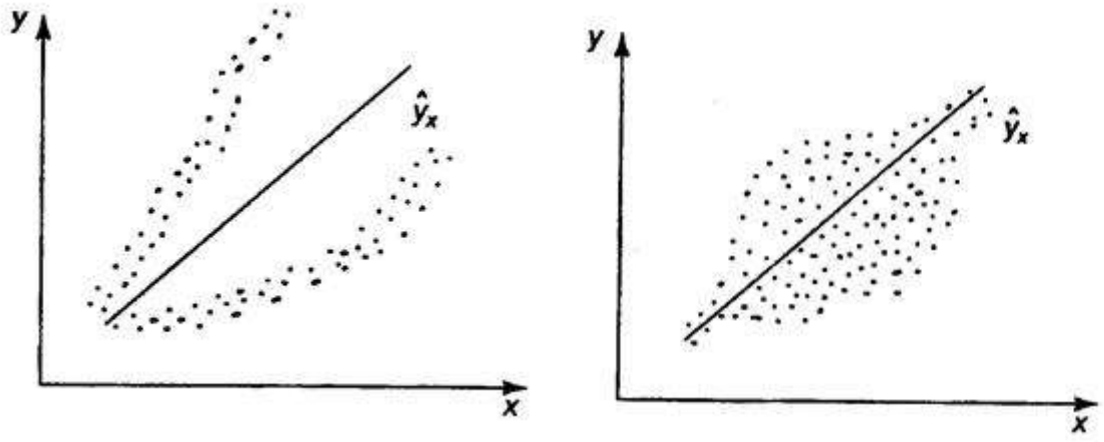


Рис. 5. Зависимость величины остатков от величины фактора x_j.

Если остатки на графике рис.5 расположены в виде горизонтальной полосы, то они независимы от значений x_j.

Если же график показывает наличие зависимости ε_j и x_j, то модель неадекватна. Причины неадекватности могут быть разные.

В соответствии с третьей предпосылкой МНК требуется, чтобы дисперсия остатков была ***гомоскедастичной.*** Это значит, что для каждого значения фактора x_j остатки ε_j имеют одинаковую дисперсию. Если это условие применения МНК не соблюдается, то имеет место ***гетероскедастичность.*** Наличие гетероскедастичности можно наглядно видеть из поля корреляции (рис. 6). Скопление точек в определенных участках значений фактора x_j говорит о наличии систематической погрешности модели.



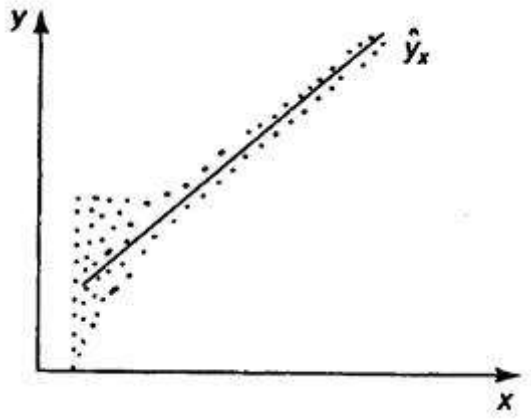


Рис.6 Примеры гетероскедастичности.

Может быть неправильна спецификация модели, и в нее необходимо ввести дополнительные члены от x_j , например x_j^2 .

Для множественной регрессии данный вид графиков является наиболее приемлемым визуальным способом изучения гомо- и гетероскедастичности.

При построении регрессионных моделей чрезвычайно важно соблюдение четвертой предпосылки МНК – отсутствие автокорреляции остатков, т.е. значения остатков ε_j, распределены независимо друг от друга. Автокорреляция остатков означает наличие корреляции между остатками текущих и предыдущих (последующих) наблюдений. Коэффициент корреляции между ε_i и ε_j, где ε_j– остатки текущих наблюдений, ε_i – остатки предыдущих наблюдений (тут $i = j - 1$), через корреляцию $K(\varepsilon_i, \varepsilon_j)$ этих остатков может быть определен как:

$$r_{\varepsilon_i \varepsilon_j} = \frac{K(\varepsilon_i, \varepsilon_j)}{\sigma_{\varepsilon_i} \sigma_{\varepsilon_j}},$$

т.е. по обычной формуле линейного коэффициента корреляции. Если этот коэффициент окажется существенно отличным от нуля, то остатки автокоррелированы и функция плотности вероятности отклонения (ошибки) $W(\varepsilon)$ зависит от предыдущей i-й точки наблюдения и от распределения значений остатков в других точках наблюдения.

При несоблюдении основных предпосылок МНК приходится корректировать модель, изменяя ее спецификацию, добавлять (исключать) некоторые факторы, преобразовывать исходные данные для того, чтобы получить оценки коэффициентов регрессии, которые

обладают свойством несмещенности, имеют меньшее значение дисперсии остатков и обеспечивают в связи с этим более эффективную статистическую проверку значимости параметров регрессии.

Обобщенный метод наименьших квадратов

При нарушении гомоскедастичности и наличии автокорреляции ошибок рекомендуется заменять традиционный метод наименьших квадратов **обобщенным методом.**

Он применяется уже к преобразованным данным и позволяет получать оценки, которые обладают не только свойством несмещенности, но и имеют меньшие выборочные дисперсии.

Например, пусть дисперсия не остается неизменной для разных значений фактора, а пропорциональна величине K_j , т.е.

$$\sigma_{\varepsilon_j}^2 = \sigma^2 K_j,$$

где $\sigma_{\varepsilon_j}^2$ – дисперсия ошибки при конкретном j-ом значении фактора; σ^2– постоянная дисперсия ошибки при соблюдении предпосылки о гомоскедастичности остатков; K_j – коэффициент пропорциональности, меняющийся с изменением величины фактора, что и обусловливает неоднородность дисперсии.

При этом предполагается, что σ^2 неизвестна, а в отношении величин K_j выдвигаются определенные гипотезы, характеризующие структуру гетероскедастичности.

В общем виде для уравнения линейной регрессии

$$y_j = a + bx_j + \varepsilon_j$$

при $\sigma_{\varepsilon_j}^2 = \sigma^2 K_j$, модель примет вид:

$$y_j = a + bx_j + \sqrt{K_j}\varepsilon_j.$$

В ней остаточные величины гетероскедастичны. Предполагая в них отсутствие автокорреляции, можно перейти к уравнению с гомоскедастичными остатками, поделив все переменные, зафиксированные в ходе j -го наблюдения, на $\sqrt{K_j}$. Тогда дисперсия остатков окажется величиной постоянной, т. е. $\sigma_{\varepsilon_j}^2 = \sigma^2$.

Иными словами, от регрессии y по x мы перейдем к регрессии на новых переменных: $y/\sqrt{K}$ и $x/\sqrt{K}$. Уравнение регрессии примет уже нормированный вид.

Применение в этом случае обобщенного МНК приводит к тому, что наблюдения с меньшими значениями преобразованных переменных $x/\sqrt{K}$ имеют при определении параметров регрессии относительно больший вес, чем с первоначальными переменными. Вместе с тем, следует иметь в виду, что новые преобразованные переменные получают новое содержание, и их регрессия имеет иной смысл, чем регрессия по исходным данным.

Таким образом обычный МНК является частным случаем обобщённого, когда весовая матрица коэффициентов w_k пропорциональна единичной. Если это не так, то преобразование в данном случае заключается в делении данных на средне-квадратичные отклонения случайных ошибок. А затем к взвешенным таким образом данным применяется обычный МНК. Как и в общем случае, дисперсии ошибок полагаются неизвестными, и их необходимо либо оценить из тех же данных.

Суммируя итог, можно сделать вывод о том, что нарушения одного или нескольких ограничений еще не приговор линейности регрессии.

1. Нелинейность регрессии может быть преодолена преобразованием переменных, например через функцию натурального логарифма ln.
2. Таким же способом возможно решить проблему неоднородной дисперсии, с помощью функций ln, или $sqrt$ преобразований зависимой переменной, либо же используя взвешенный МНК.
3. Для устранения проблемы мультиколлинеарности применяется метод исключения переменных. Суть его в том, что высоко коррелированные объясняющие переменные устраняются из регрессии, и она заново оценивается. Критерием отбора переменных, подлежащих исключению, является коэффициент корреляции. Есть еще один способ решения данной проблемы, который заключается в замене переменных, которым присуща мультиколлинеарность, их линейной комбинацией.

К сожалению, не все нарушения условий и дефекты линейной регрессии можно устранить с помощью натурального логарифма. Если

имеет место *автокорреляция возмущений* к примеру, то лучше отступить на шаг назад и построить новую и лучшую модель.

Этим весь список не исчерпывается, есть еще ***пошаговая рекурсивная регрессия*** **и другие методы**.

Рекурсивный (рекуррентный) метод наименьших квадратов

Рассмотренный нами метод наименьших квадратов требует для своей реализации хранения большого объема данных и не всегда применим в реальных условиях при обработке очень больших массивов данных. Кроме того часто по смыслу решаемой задачи требуется **последовательная обработка вновь поступающих наблюдений** и необходимо принимать решение на основе вновь поступающей информации. В этом случае применяют рекуррентную форму метода наименьших квадратов.

Предположим, что уже получены оценки ее коэффициентов $\overline{b_{n-1}}$ по данным выборочной совокупности из $(n-1)$ наблюдения. Требуется в ситуации, когда в выборочную совокупность добавлено новое наблюдение (y_n, x_n), пересчитать оценки коэффициентов регрессии, используя для этого ранее полученные оценки $\overline{b_{n-1}}$.

$$\overline{b_n} = \overline{b_{n-1}} + \frac{\overline{b_{n-1}}^{-1} x_n'}{\left(x_n \overline{b_{n-1}}^{-1} x_n' + 1\right)} \left[y_n - x_n \overline{b_{n-1}}\right]$$

Полученная формула позволяет осуществлять пересчет оценок рекуррентно по мере появления новых наблюдений. С ее помощью реализуются основные идеи построения адаптивных многофакторных регрессионных моделей.

Несмотря на то, что параметры регрессионной модели, как правило, оцениваются с использованием метода наименьших квадратов, существуют и другие методы, которые используются гораздо реже. К примеру, это следующие методы:

- Байесовские методы (например, байесовский метод линейной регрессии).
- Процентная регрессия, использующаяся для ситуаций, когда снижение процентных ошибок считается более целесообразным.

- Наименьшие абсолютные отклонения, что является более устойчивым в присутствии выбросов, приводящих к квантильной регрессии.
- Непараметрическая регрессия, требующая большого количества наблюдений и вычислений. Расстояние метрики обучения, которая изучается в поисках значимого расстояния метрики в заданном входном пространстве.

Сглаживание функций сплайнами.

При использовании сглаживания (аппроксимации) функций меру их близости к оригиналу можно определять по-разному, то это приводит к значительному разнообразию сглаживающих функций и появлению, например, сглаживающих сплайнов. И если первоначально рассматривались только кусочно-полиномиальные сплайны третьей и иногда выше степеней, то по мере расширения сферы их приложений, стали возникать сплайны «склеенные» из других элементарных функций.

Также, как и в интерполяционном сплайне описывается гладкая, дважды дифференцируемая функция $g(x)$ в виде полинома третьей степени. Определяются краевые условия. Отличие правил построения сглаживающего сплайна от интерполяционного состоит в том, что пользователю необходимо определить некоторые весовые коэффициенты, которые обеспечивают минимум функционалу

$$\Phi(g) = \int_a^b \left(g''(x)\right)^2 dx + \sum_{i=0}^{n} (g(x_i) - z_i)^2 / r_i \ ,$$

так, что $|g(x_i) - z_i| \le \delta_i$, $0 \le i \le n$, δ_i – некоторые числа, $z_i = y_i + \varepsilon_i$, ε_i- погрешность измерения, r_i - заданные положительные числа, называемые весовыми коэффициентами. Весовые коэффициенты r_i задаваемые пользователем, позволяют в известной степени управлять свойствами сглаживающих сплайнов. Если все $r_i = 0$, то $z_i = y_i$, и сглаживающий сплайн становится интерполяционным.

Таким образом, интерполяционный кубический сплайн можно рассматривать как частный случай сглаживающего кубического сплайна для всего набора измерений $0 \le i \le n$. Отметим, что чем меньше погрешность ε_i, тем меньше должны быть весовые коэффициенты r_i . Если же необходимо, чтобы сглаживающий

кубический сплайн прошел через некоторую точку (x_k, z_k), то соответствующий весовой множитель r_kследует положить равным нулю. В практических вычислениях выбор величин r_i при постановке задачи моделирования является важным вопросом для решения которого используются различные подходы.

Сложность вычислительного алгоритма построения сглаживающего кубического сплайна та же, что и у интерполяционного кубического сплайна. Число арифметических действий, необходимых для построения сглаживающего кубического сплайна, пропорционально числу отрезков $(O(n))$. Объем вычислительной памяти также пропорционален числу отрезков $(O(n))$.