Тема «Регрессионный анализ»

Выведение линии тренда

Линии тренда — популярный инструмент для прогнозирования, поскольку они просты как для вычисления, так и для понимания. Достаточно открыть любую ежедневную газету, чтобы увидеть графики трендов в самых различных областях: от цен на акции до прогноза погоды. Общие тренды обычно применяют единственный предиктор для предсказания результата, используя, например, время (предиктор) для прогнозирования цен на акции компании (результат). Однако можно улучшить предсказание цен на акции, добавив другие предикторы, такие как уровень продаж. Это становится возможным с регрессионным анализом, позволяющим не только улучшать прогнозирование путем учета множества предикторов, но и сравнивать эти предикторы между собой по степени влияния. Чтобы разобраться с этим, посмотрим на пример с предсказанием цен на дома.

6.2. Пример: предсказание цен на дома.

Здесь использовали данные за 1970-е годы о ценах на дома в Бостоне. Предварительный анализ показывает, что двумя сильнейшими предикторами цен на дома являются число комнат в доме и доля соседей с низким доходом. На рис. 1 видно, что у дорогих домов обычно больше комнат.

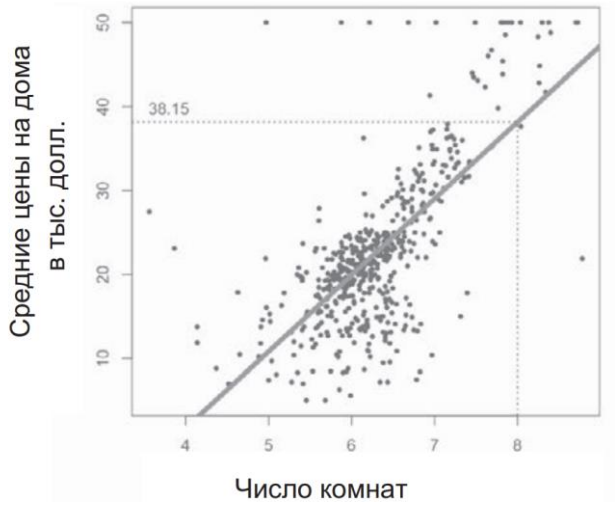


Рис. 1. Цены на дома в сравнении с числом комнат

Для предсказания цены дома можно построить линию тренда, известную также как линия наилучшего соответствия. Она проходит близко к наибольшему числу элементов данных на графике. Например, если у дома восемь комнат, его цена составит приблизительно $ 38 150.

Кроме числа комнат на цену дома также влияло его окружение. Дома оказались дешевле там, где была выше пропорция соседей с низким доходом (рис. 2). Поскольку тренд получался немного изогнутым (рис. 2, а), к предикторам применили математическую

операцию, известную как взятие логарифма. Благодаря этому через элементы данных проще провести прямую линию тренда (рис. 2, b). Можно заметить, что элементы данных на рис. 2, b плотнее прилегают к линии тренда, чем на рис. 1. Это означает, что фактор соседства оказался более точным предиктором цены дома, чем число комнат.

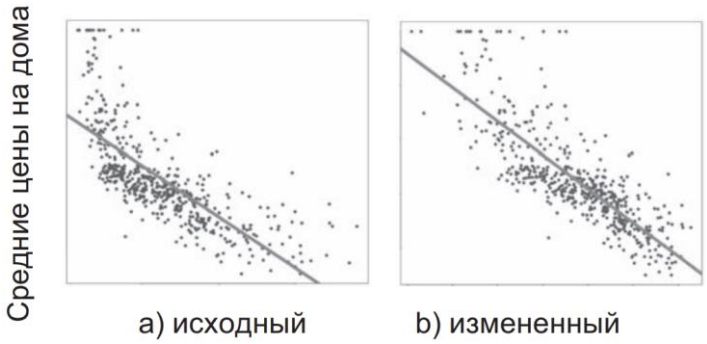


Рис.2. Цены на дома в сравнении с долей соседей с низким доходом

Для улучшения наших расчетов цен на дома мы можем учесть и число комнат, и влияние соседства. Но поскольку выяснилось, что влияние соседства лучше предсказывает цену дома, простое сложение этих двух предикторов не станет идеальным решением. Вместо этого предиктору соседства нужно задать больший вес.

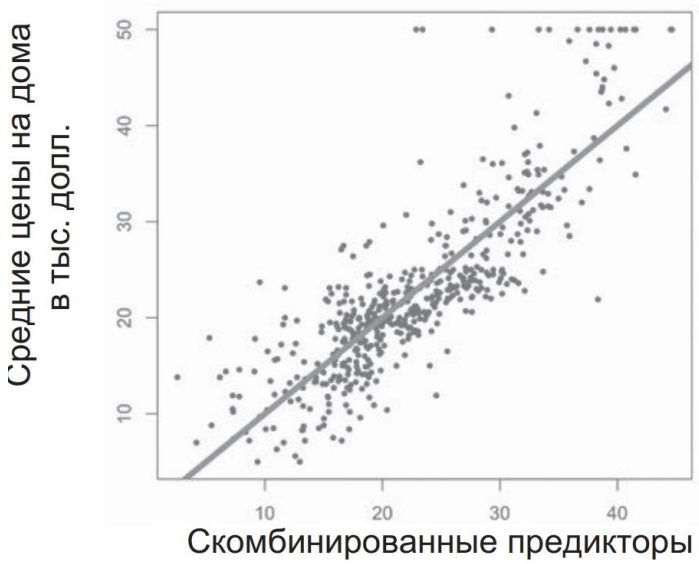


Рис. 3. Цены на дома в сравнении со скомбинированным предиктором из числа комнат и доли соседей с низким доходом

Рис. 3 показывает график цен на дома согласно оптимальной комбинации двух предикторов. Обратите внимание на то, что элементы данных располагаются еще ближе к итоговой линии тренда, чем раньше, поэтому прогноз с использованием такой линии тренда должен оказаться точнее. Чтобы проверить это, можно сравнить погрешность трех линий тренда (табл. 1).

Таблица 1. Средняя прогностическая ошибка при использовании трех разных линий тренда

	Погрешность прогнозирования (в тыс. долл.)
Число комнат	4,4
Влияние окружения	3,9
Число комнат и влияние окружения	3,7

Хотя очевидно, что уравновешенная комбинация предикторов ведет к более точным предсказаниям, возникают два вопроса:

1) как вычислить оптимальный вес предикторов;

2) как следует их проинтерпретировать.

Градиентный спуск

Вес предиктора — главный параметр регрессионного анализа, и оптимальный вес обычно вычисляется путем решения уравнений. Тем не менее, поскольку регрессионный анализ прост и годится для визуализации, воспользуемся им для демонстрации альтернативного способа оптимизации параметров. Этот метод называется градиентным спуском (gradient descent) и используется в случаях, когда параметры нельзя получить напрямую.

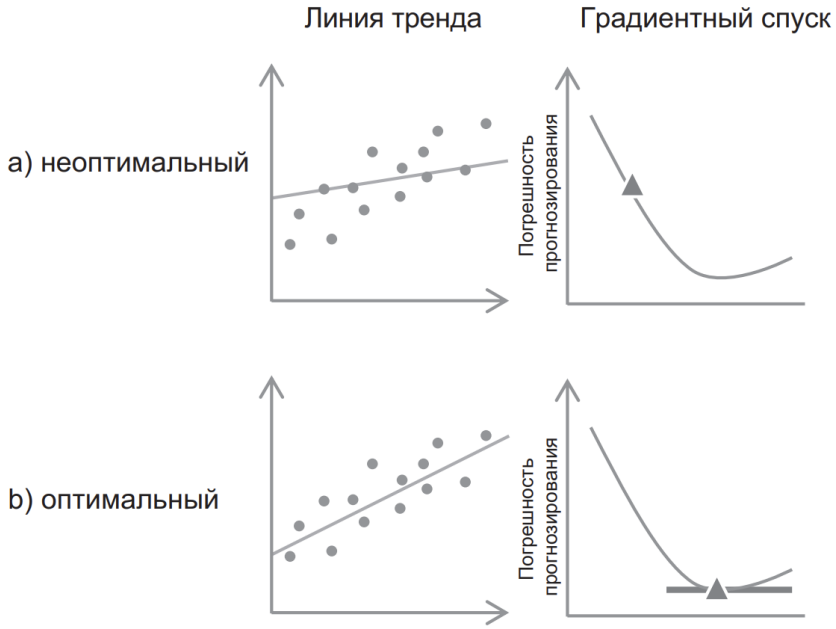


Рис. 4. Как линия тренда достигает оптимальности благодаря градиентному спуску

Вкратце: алгоритм градиентного спуска делает первоначальное предположение о наборе весовых составляющих, после чего начинается итеративный процесс их применения к каждому элементу данных для прогнозирования, а затем они перенастраиваются для

снижения общей ошибки прогнозирования. Этот процесс можно сравнивать с пошаговым спуском в овраг в поисках дна. На каждом этапе алгоритм определяет, какое направление даст наиболее крутой спуск, и пересчитывает весовые составляющие. В конечном итоге мы достигнем самой нижней позиции, которая представляет собой точку, в которой погрешность прогнозирования минимальна. Рисунок 4 показывает, как оптимальная линия тренда регрессии соответствует нижней точки градиента.

Кроме регрессии градиентный спуск может также использоваться для оптимизации параметров в других моделях, таких как метод опорных векторов или в нейронных сетях. Однако в этих более сложных моделях результаты градиентного спуска могут зависеть от стартовой позиции в овраге (то есть изначальных значений параметра). Например, если нам случится начать в небольшой яме, алгоритм градиентного спуска может ошибочно принять это за оптимальную точку (рис. 5).

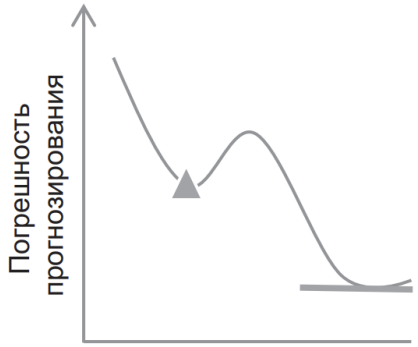


Рис. 5. Как ближайшая яма может быть ошибочно принята за оптимальную точку (треугольник), хотя истинная оптимальная точка находится ниже ее (черта)

Чтобы снизить риск попадания в такую яму, можно воспользоваться стохастическим градиентным спуском, при котором вместо использования всех элементов данных для регулировки параметров при каждой итерации берется только один. Это привносит вариативность, позволяя алгоритму избегать ям. Хотя итоговые значения параметров после работы стохастического процесса могут оказаться не оптимальными, они, как правило, обеспечивают достаточно высокую точность. Тем не менее этот «недостаток» относится только к более сложным моделям, и нам не о чем беспокоиться, когда мы используем регрессионный анализ.

Коэффициенты регрессии

После получения оптимального набора регрессионных предикторов их нужно интерпретировать. Вес регрессионных предикторов называется коэффициентом регрессии. Коэффициент регрессии показывает то, насколько силен предиктор при совместном использовании с другими. Иными словами, это значение, добавляемое к предиктору, а не его собственная предсказательная способность. Например, если кроме числа комнат

использовать для предсказания цены дома его общую площадь, то значимость числа комнат может показаться незначительной. Поскольку и число комнат, и общая площадь дома связаны с его размером, это добавляет к предсказательной силе не так уж и много. Толковой интерпретации регрессионных коэффициентов мешает также различие в единицах измерения. Например, если предиктор измеряется в сантиметрах, его вес будет в 100 раз отличаться по весу от предиктора, берущегося в метрах. Чтобы избежать такого, нужно стандартизировать единицы измерения предикторных переменных перед тем, как проводить регрессионный анализ.

Стандартизация — это выражение переменных в процентилях. Когда предикторы стандартизированы, то коэффициент, который называется бета-весом, может быть использован для более точных сравнений. В примере с ценами на дома два предиктора (первый — число комнат, второй — соседи с низким доходом) были стандартизированы в соотношении 2,7 к 6,3. Это означает, что доля жильцов с низким доходом является более мощным предиктором цены на дом, чем количество комнат. Уравнение регрессии будет выглядеть примерно так: *цена = 2,7 (количество комнат) – 6,3(% соседей с низким доходом)*. Обратите внимание, что в этом уравнении доля жильцов с низким доходом имеет отрицательный вес, что выражено знаком «минус». Дело в том, что предиктор имеет обратную корреляцию с ценами на дома, как показано на устремленной вниз линии тренда на рис. 2.

Коэффициенты корреляции

Если предиктор только один, бета-вес такого предиктора называется коэффициентом корреляции и обозначается как *r*. Коэффициенты корреляции варьируются от –1 до 1 и несут две единицы информации.

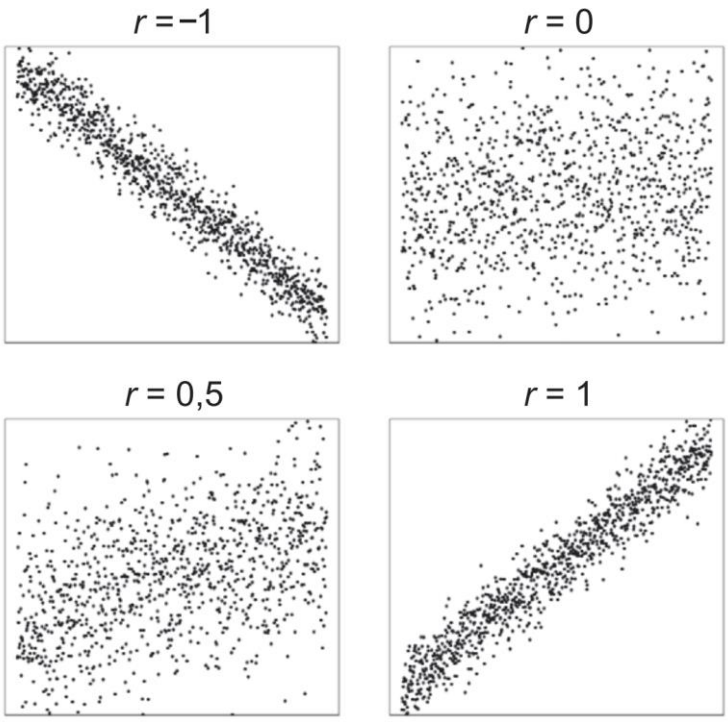


Рис. 6. Пример распределения данных в соответствии с различными коэффициентами корреляции

Направление. При положительных коэффициентах предиктор стремится в том же направлении, что и результат. При отрицательных — в обратном направлении. Цены домов положительно коррелируют с числом комнат, но отрицательно коррелируют с долей жильцов с низким доходом по соседству.

Величина. Чем ближе коэффициент к −1 или 1, тем сильнее предиктор. Например, коэффициент корреляции, показанный линией тренда на рис. 1, равен 0,7, в то время как на рис. 2, b это –0,8. Это означает, что достаток соседей — более достоверный предиктор цен на дома, чем число комнат. Нулевая корреляция означала бы отсутствие связи между предиктором и результатом. Коэффициенты корреляции показывают абсолютную силу отдельных предикторов и, следовательно, являются более надежным способом их ранжирования, чем коэффициенты регрессии.

Ограничения

Несмотря на то что регрессионный анализ информативен и не требует долгих вычислений, он имеет недостатки.

Чувствительность к резко отклоняющимся значениям. Регрессионный анализ одинаково учитывает все предоставленные элементы данных. Если среди них будет хотя бы несколько элементов с крайними значениями, это может значительно исказить линию тренда. Чтобы избежать этого, можно использовать диаграмму рассеяния для предварительного выявления таких резко отклоняющихся значений.

Искажение веса при корреляции предикторов. Включение в регрессионную модель высококоррелирующих предикторов исказит интерпретацию их веса. Эта проблема называется мультиколлинеарностью. Для преодоления мультиколлинеарности нужно либо исключить из анализа коррелирующие предикторы, либо воспользоваться более продвинутым методом, таким как лассо или риджрегрессия (или гребневая регрессия).

Криволинейные тренды. В нашем примере тренды отображались прямой линией. Тем не менее некоторые тренды могут быть криволинейными, как на рис. 2, а. В этом случае нам потребуется преобразовать значения предикторов или использовать альтернативные алгоритмы, такие как метод опорных векторов.

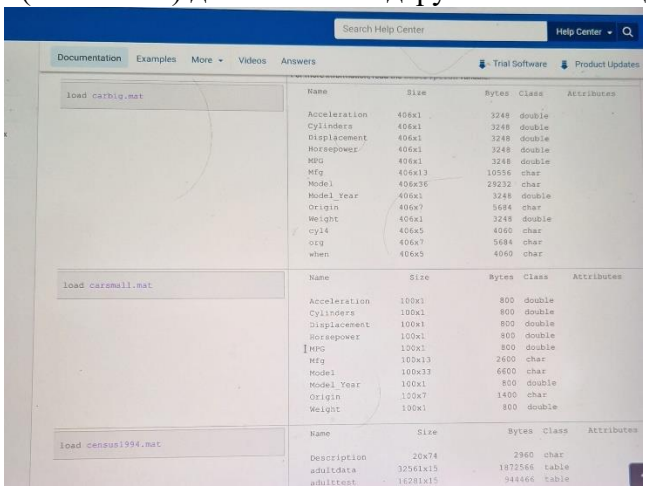
Корреляция не говорит о причинности. Предположим, была обнаружена положительная корреляция между стоимостью дома и наличием собаки. Понятно, что если просто завести собаку, цена дома от этого не изменится, однако можно предположить, что те, кто могут позволить себе содержать собак, располагают в среднем большим доходом и, вероятно, проживают в районах, где дома стоят дороже.

Несмотря на эти ограничения, регрессионный анализ остается одним из основных, простых в использовании и интуитивно-понятных методов для прогнозирования. Внимательное отношение к способу интерпретации результатов — залог уверенности в точности выводов.

Лабораторная работа 5.

Задача для примера:

Необходимо построить линию регрессию для набора данных. Исходный набор данных (`carsmall`) для MatLab подгружен из имеющегося dataset



Здесь собраны данные по машинам за разные годы, ускорение, количество цилиндров, перемещение, мощность в л.с., расход (MPG - сколько можно проехать на одном галоне), модель, год создания модели, масса машины и пр.

Далее можно проводить анализ, строить зависимости и строить линии тренда

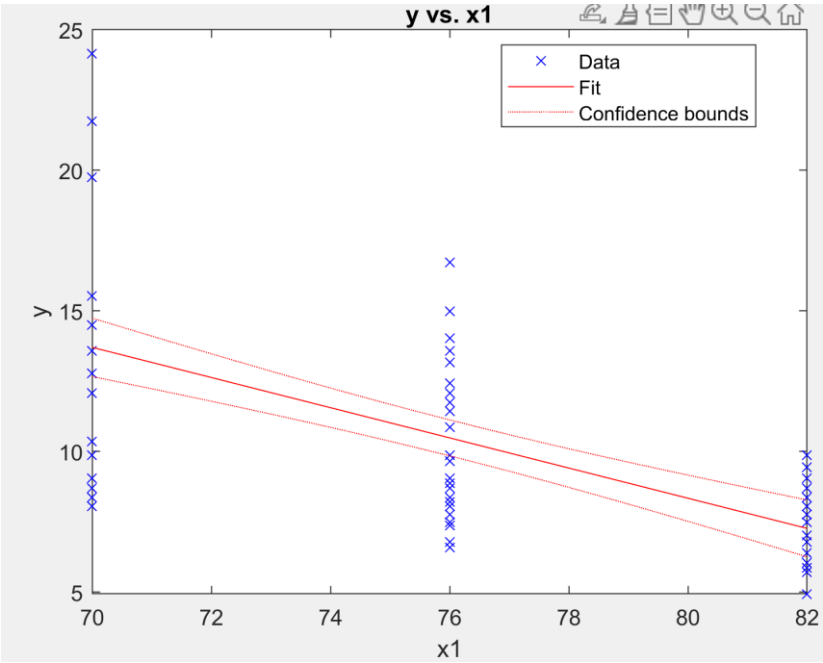
Пример кода

```
load carsmall
X = [Model_Year];
lp100km=1./(MPG*0.46)*100;
mdl = fitlm(X,lp100km);

figure(50)
plot(mdl);
hold on;

x1 =[70,76,82,85];
y1=-65.703+1.1746*x1;
plot(x1,y1,'*-');
```

В данном примере строим линейную регрессиию `lp100km` от года выпуска машины.



где y – это литров на 100 км, x1 – это год создания модели.

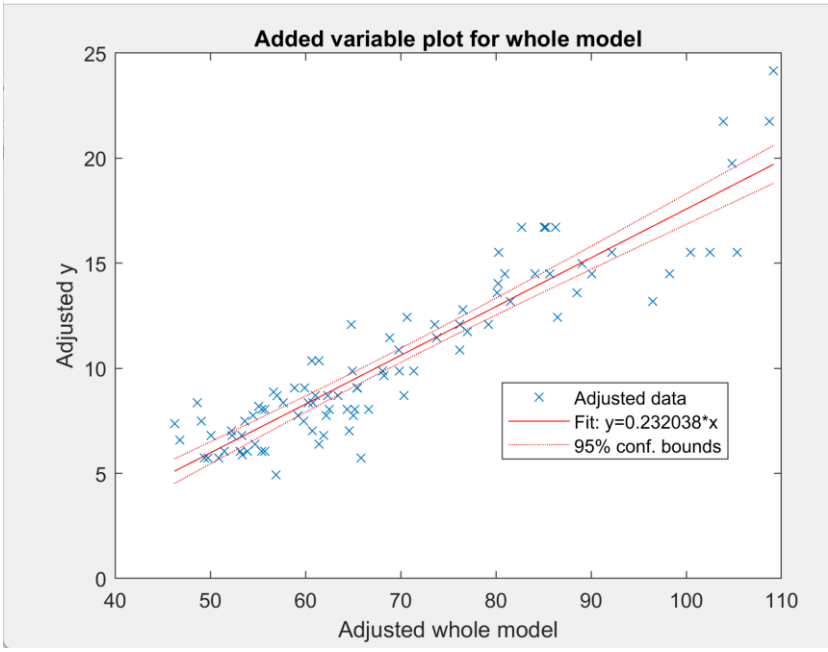
Построена линия тренда и доверительный диапазон. Как можно видеть, линия тренда показывает, что в 1970 году стандартно тратилось примерно 14 литров на 100 км, а в 1982 год у появились автомобили с расходом топлива 8 литров на 100 км. И виден тренд, который показывает снижение количества литров на 100 км.

Можно строить линию тренда по группе данных (данные берем осмысленно).

Выбрали вес, лошадиные силы и ускорение. Т.е. за сколько разгоняется автомобиль до 100 км\ч.

```
X = [Weight,Horsepower,Acceleration];
```

Если используется много факторов, то алгоритм выбирает веса под каждый фактор (подбирает нормировочный коэффициент так, чтобы можно было построить линейную регрессию).



Снова видна четная линия тренда и доверительный диапазон (в нашем случае доверительная вероятность составила 95%).

И получили осредненную функцию для линии тренда *y=0,232038*x.*

Задание для лабораторной работы:

Построить линейную регрессию и линию тренда для любого интересующего набора данных.