ЦИФРОВАЯ КУЛЬТУРА
УНИВЕРСИТЕТ ИТМО

Регрессия

Высшая Школа Цифровой Культуры
Университет ИТМО
dc@itmo.ru

Содержание

1 Простейшая линейная регрессия

Итак, мы приступаем к решению первой задачи обучения с учителем – задаче регрессии. Как уже отмечалось, задача регрессии – это задача предсказания числа (или отклика) Y по значениям входных переменных $X_1, X_2, ..., X_p$ (или предикторов). Функцию $f(X)$, отвечающую зависимости $Y = f(X_1, X_2, ..., X_p)$, мы будем предполагать линейной, а наша задача будет заключаться в поиске коэффициентов этой линейной модели. Говоря математическим языком, мы будем решать задачу параметрического оценивания. Начнем?

1.1 Линейная регрессия и МО

Часто требуется определить, как зависит одна случайная величина от одной или нескольких других величин. Самый общий вид зависимости – статистическая зависимость. Например, пусть $X = \xi + \eta$ – это сумма случайных величин ξ и η, а $Y = \xi + \varphi$ – сумма случайных величин ξ и φ. Ясно, что величины X и Y зависимы, но нет явной функциональной зависимости, то есть мы не можем указать зависимость вида $X = f(Y)$ или $Y = f(X)$.

Можно дать и более неформальный и жизненный пример. Ясно, что стоимость квартиры зависит от площади, этажа, месторасположения и других параметров, но не является функцией от них. Все потому, что есть куча факторов, которые просто невозможно учесть. Например, при одинаковых входных параметрах (хотя и это очень относительно) продавец, скорее всего, выставит квартиру дешевле, если ему срочно нужны деньги, и не будет снижать цену ни на рубль, если продажа «не горит», а может и вообще поднять ее из-за того, что каждую весну на балконе ласточки вьют гнездо. Ну и как тут понять ценообразование?

Что же в этом случае делать? Как получить хоть какую-то функцию, которая может предсказать изменение интересующей нас величины по изменению параметров? Для зависимых случайных величин имеет смысл рассмотреть математическое ожидание одной из них при фиксированном значении другой и выяснить, как влияет на среднее значение первой величины изменение значений второй. Так, в примере с квартирой, среднее значение цены можно считать функцией от параметров, влияющих на цену.

В этой части мы познакомимся с понятием линейной регрессии, достаточно простым и часто используемым «инструментом» при обучении с учителем. Линейная регрессия известна уже довольно давно и подробно освещена в большом количестве книг. На первый взгляд может показаться, что она слишком тривиальна по сравнению с более продвинутыми средствами статистики, о которых будет рассказано позже, но на самом деле линейная регрессия до сих пор широко применяется как непосредственно в статистике, так и в ее

приложениях к машинному обучению. Кроме того, линейная регрессия является хорошей отправной точкой для изучения более новых подходов, так как многие методы статистики, как мы увидим позже, есть не что иное, как обобщение линейной регрессии.

На какие же вопросы может ответить линейная регрессия? Для иллюстрации приведем пример. Предположим, что мы – консультанты-аналитики, работающие в некоторой фирме, перед которыми стоит задача анализа и улучшения объема продаж определенного продукта. Пусть в качестве продуктов выступают, например: мобильные телефоны и самолеты. Эти продукты продаются у ста одного дистрибьютора (объем выборки – 101). В качестве входных данных выступает объем финансирования, вложенного в рекламу конкретного продукта (в тысячах и сотнях тысячах долларов), а в качестве выходных – объем проданного товара (в тысячах единиц). Еще раз поясним, что на рисунках по оси абсцисс отложено количество финансирования, выделенного на рекламу продукции, а на оси ординат – объем продаж продукта (в тысячах единиц). Рисунок 1 отвечает за мобильные телефоны, а рисунок 2 – за самолеты. Уже на первом рисунке мы видим, что реклама, в общем и целом, продуктивно влияет на объем продаж телефонов, хотя вид зависимости не очень понятен.

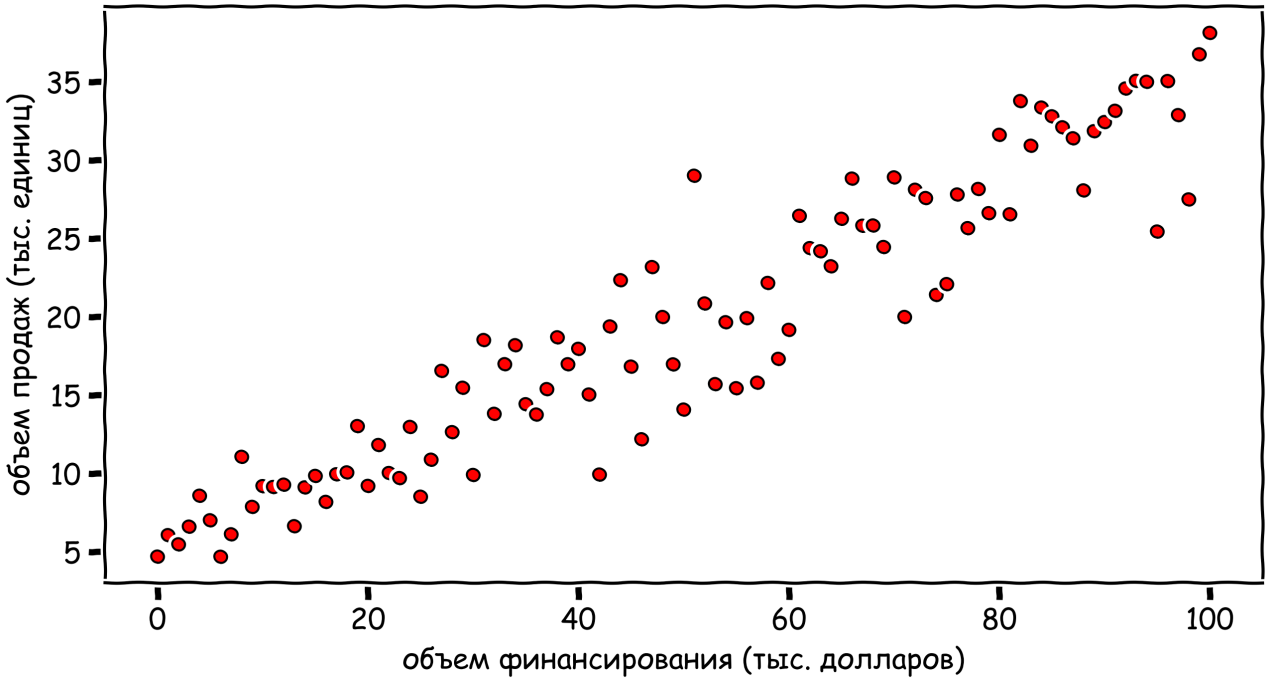

Рис. 1: Зависимость объема продаж мобильных телефонов от затрат на рекламу

Совершенно наивно полагать, что самая «правильная» зависимость – это зависимость, представленная на рисунке 3 (зависимость получена просто соединением соседних точек отрезками). Как интерпретировать такую модель,

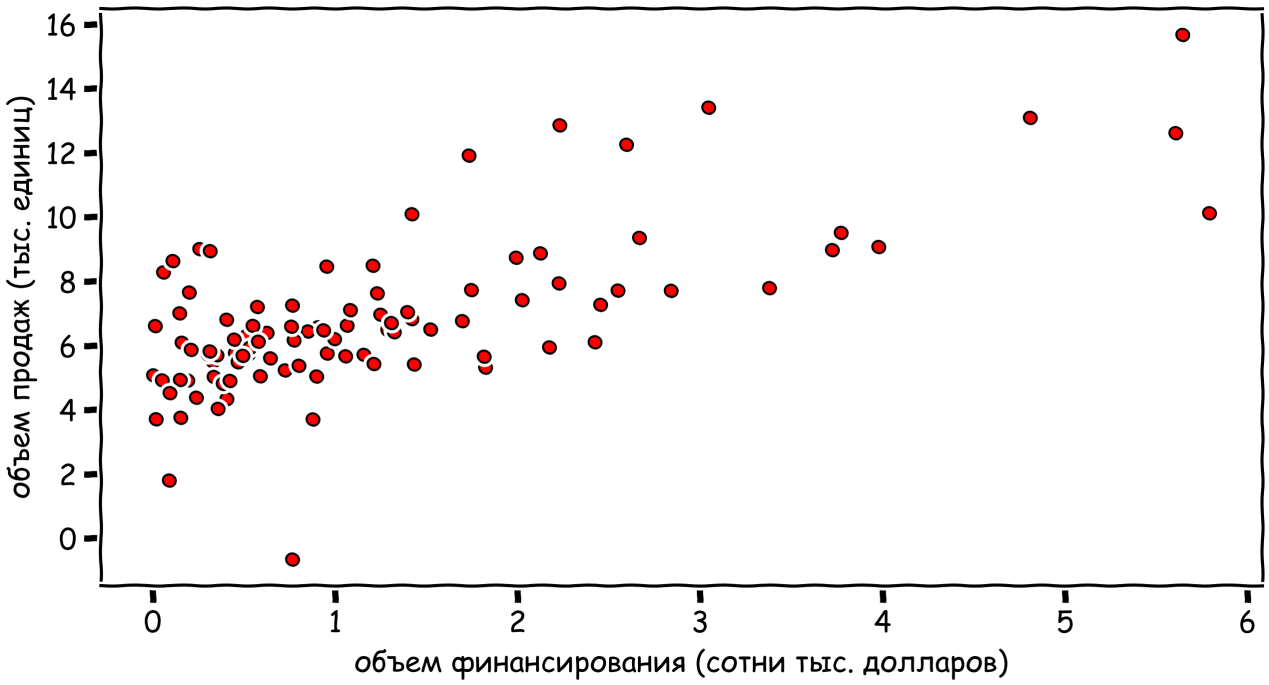

Рис. 2: Зависимость продаж самолетов от объема финансирования

как объяснить? Почему при увеличениях затрат на рекламу продажи то резко падают, то взмывают вверх? Может быть есть какие-то неучитываемые нами параметры, как, например, период отпусков (когда продажи падают по объективным причинам), или приближение нового года (когда они же взмывают вверх, и снова понятно почему), а может данные просто содержат ошибки? Во всех этих случаях предложенная «модель» только усугубит прогноз и будет ни чем не лучше, чем просто число, сказанное наугад.

Второй рисунок трактовать сложнее. Мы видим, что небольшое финансирование (до 100 тысяч долларов), в общем и целом, дает примерно одинаковый объем продаж, хотя имеются и выбросы в сторону увеличения объема. Дальше же ситуация противоречива. Увеличение затрат на рекламу до 400, а то и до 600 тысяч долларов в среднем увеличивает продажи в полтора-два раза, опять же, за исключением некоторых выбросов. Кстати, на втором рисунке видны и очевидно аномальные данные с отрицательным объемом продаж.

Директор фирмы не может непосредственно повлиять на объем продаж, однако он может влиять на объем бюджета, выделяемого на рекламу, косвенно влияя на продажи. Какие же вопросы нас могут заинтересовать?

1. Есть ли реальная зависимость между вложенным в рекламу бюджетом и объемом продаж? Ясно, что если зависимости не наблюдается, то зачем тратить деньги на рекламу?

2. Если зависимость все-таки есть, то насколько она сильна? Другими сло-

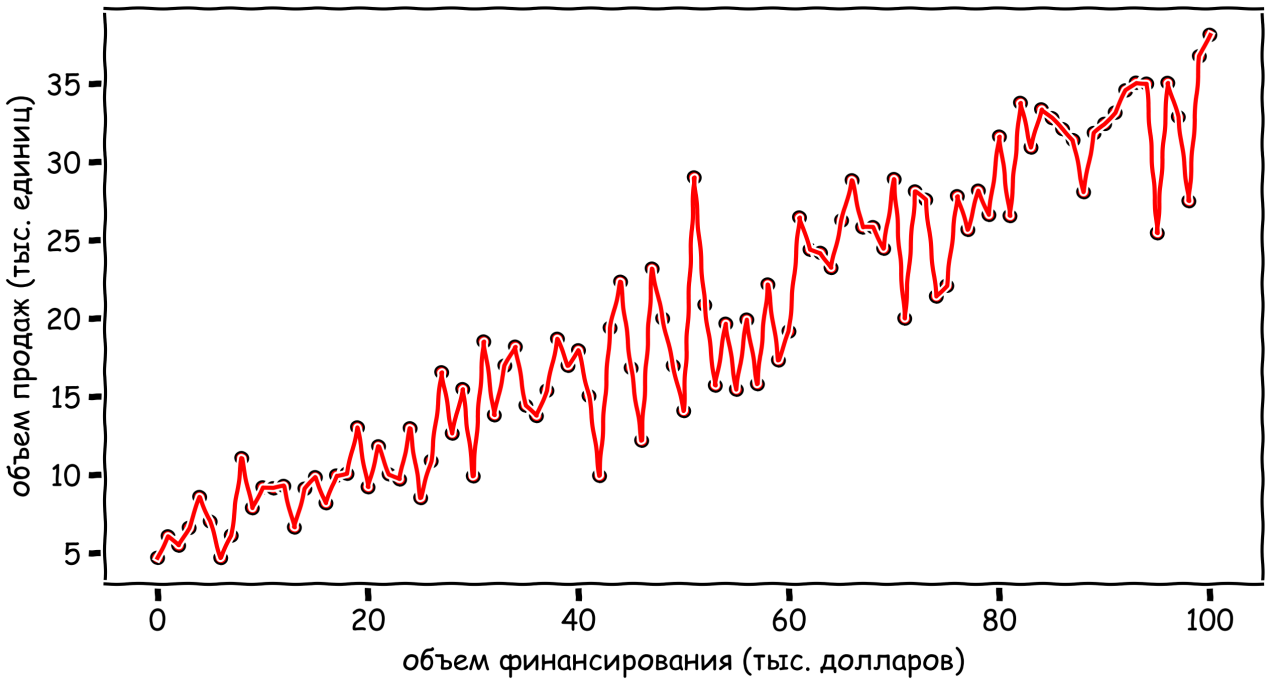

Рис. 3: Наивная зависимость

вами, зная объем бюджета, потраченного на рекламу, можем ли мы с достаточной точностью предсказать объем продаж? Если да, то зависимость сильная, иначе – слабая.

3. Какие товары популярны и продаются? Выгодно ли тратить бюджет на рекламу всех товаров?

4. Насколько точно мы можем оценить изменение объема продаж, изменяя объем бюджета для рекламы?

5. Насколько точно мы можем предсказать объем продаж, зная объем вливаемого в рекламу бюджета?

6. Линейна ли зависимость?

7. Имеется ли синергия в областях продаж? Ведь может так оказаться, что вливание 50000 долларов на рекламу мобильных телефонов и 50000 долларов на рекламу самолетов лучше, то есть приведет к более высокому объему продаж, чем вливание 100000 на рекламу только телефонов.

Оказывается, линейная регрессия может ответить на каждый из написанных выше вопросов. Давайте приступим к детальному изучению.

1.2 Модель простейшей линейной регрессии и метод наименьших квадратов

Предположим, что наблюдаемая (случайная) величина Y зависит от некоторого известного и неслучайного фактора X_1, а также от случайной ошибки ε, наличие которой объясняется либо погрешностью измерений, либо ошибками самой модели, либо эта ошибка просто-напросто заложена в основе эксперимента. В качестве основной модели в этом пункте мы будем рассматривать следующую линейную модель

$$Y = \theta_0 + \theta_1 X_1 + \varepsilon.$$

Итак, зависимость между Y и X_1 предполагается линейной с точностью до некоторой ошибки.

Определение 1.2.1 *Модель, описываемая зависимостью*

$$Y = \theta_0 + \theta_1 X_1 + \varepsilon,$$

где θ_0, θ_1 – числовые параметры, X_1 – неслучайный параметр, значения которого либо задаются, либо наблюдаются (иначе говоря – известны), ε – случайная ошибка, называется моделью простейшей линейной регрессии.

Часто используют и следующие два определения.

Определение 1.2.2 *Функция*

$$f(X_1) = \theta_0 + \theta_1 X_1$$

в модели простейшей линейной регрессии называется линией регрессии Y на X_1.

Определение 1.2.3 *Уравнение*

$$Y = \theta_0 + \theta_1 X_1$$

в модели простейшей линейной регрессии называется уравнением регрессии Y на X_1.

Абстрактная модель – это хорошо. Но как ее строить и применять на практике, на конкретных наблюдаемых значениях X_1 и Y? Давайте опишем схему подробнее. Начнем же с того, что аккуратно выпишем: а что дано?

Итак, пусть проводится n экспериментов, в каждом из которых (обратите на это внимание!) **неслучайная** величина X_1 приняла значения $x_1, x_2, ..., x_n$.

Допустим также, что среди этих значений есть хотя бы два различных. В зависимости от значений величины X_1, мы наблюдаем n значений $Y_1, Y_2, ..., Y_n$ нашей **случайной** величины Y. В итоге, в результате эксперимента у нас есть n пар данных $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$. Ясно, что они могут быть легко изображены на плоскости. На рисунке 4, показывающем зависимость объема продаж мобильных телефонов в зависимости от затрат на их рекламу, изображена 101 пара данных. **Неслучайная** величина X_1 – это **известный** объем выделенного финансирования рекламы, а **случайная** величина Y – это объем продаж при известном финансировании и **неизвестных, случайных** других факторах (или ошибках).

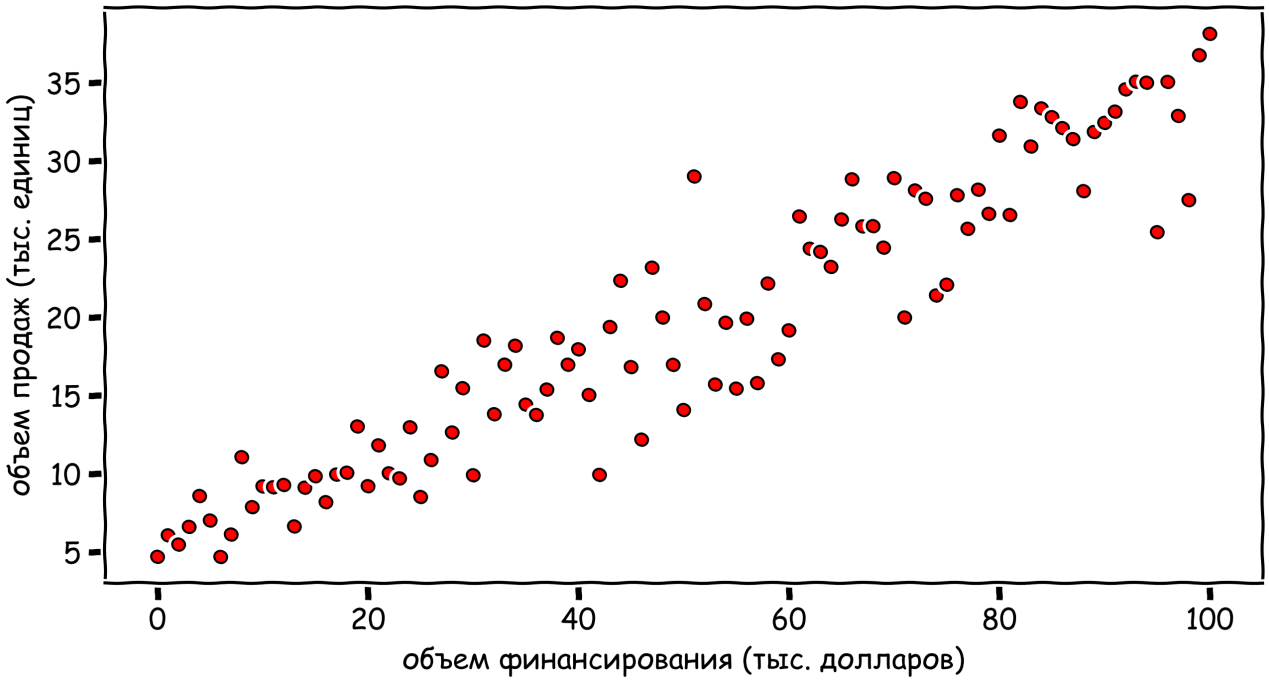

Рис. 4: Зависимость объема продаж мобильных телефонов от затрат на рекламу

В итоге, так как в результате измерений в эксперименте возникали случайные ошибки, о возможной природе которых мы говорили ранее, то точного равенства

$$Y_i = \theta_0 + \theta_1 x_i$$

для каждого измерения $i \in \{1, 2, ..., n\}$ при одних и тех же (пока что неизвестных!) параметрах θ_0 и θ_1 получить, скорее всего, не получится, но можно утверждать, что при каждом i справедливо соотношение

$$Y_i = \theta_0 + \theta_1 x_i + \varepsilon_i$$

при одних и тех же параметрах θ_0 и θ_1.

Замечание 1.2.1 *Отметим в качестве замечания, что конкретные значения $Y_1, Y_2, ..., Y_n$ случайной величины Y вполне разумно обозначать маленькими буквами $y_1, y_2, ..., y_n$. Однако во избежании путаницы с тем, что Y случайна, а X_1 – нет, договоримся в этой лекции и конкретные значения случайной величины Y обозначать заглавными буквами.*

Итак, сама по себе модель озвучена. Но как же найти оценки параметров θ_0 и θ_1, зная набор данных $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$? Ведь не зная параметров, мы не можем решать задачу предсказания (а именно для ее решения все и затевается).

Для поиска коэффициентов регрессии, мы будем пользоваться методом наименьших квадратов (МНК). Этот метод позволяет найти такие оценки $\widehat{\theta}_0$ и $\widehat{\theta}_1$ параметров θ_0 и θ_1, что сумма квадратов ошибок $\varepsilon(\theta_0, \theta_1)$ в наблюдаемых n экспериментах минимальна. Иными словами, минимизируется функция

$$\varepsilon(\theta_0, \theta_1) = \varepsilon_1^2 + \varepsilon_2^2 + ... + \varepsilon_n^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1 x_i)^2$$

и ищутся аргументы, ее минимизирующие. В итоге решается следующая задача

$$\arg\min_{\theta_0, \theta_1} \sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1 x_i)^2 .$$

Определение 1.2.4 *Оценкой метода наименьших квадратов для неизвестных параметров θ_0 и θ_1 уравнения регрессии называется набор значений параметров, минимизирующий выражение*

$$\varepsilon(\theta_0, \theta_1) = \sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1 x_i)^2 .$$

Ну что, идейная сторона вопроса на этом закончена. С технической же точки зрения перед нами функция

$$\varepsilon(\theta_0, \theta_1) = \sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1 x_i)^2 ,$$

зависящая от двух переменных θ_0 и θ_1, которую нам требуется минимизировать. Решение задачи минимизации дается следующей теоремой.

Теорема 1.2.1 *Минимум функции*

$$\varepsilon(\theta_0, \theta_1) = \sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1 x_i)^2$$

единственен и достигается при

$$\theta_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})(Y_i - \overline{Y})}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}, \quad \theta_0 = \overline{Y} - \theta_1\overline{X_1},$$

где $\overline{X_1}$ *– среднее принимаемых переменной* X_1 *значений, то есть*

$$\overline{X_1} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

а $\overline{Y}$ *– среднее принимаемых переменной* Y *значений, то есть*

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Доказательство. Рассматриваемая функция дифференцируемая, а значит необходимым условием экстремума является равенство нулю частных производных этой функции:

$$\begin{cases} \frac{\partial\varepsilon(\theta_0,\theta_1)}{\partial\theta_0} = 0 \\ \frac{\partial\varepsilon(\theta_0,\theta_1)}{\partial\theta_1} = 0 \end{cases} \Leftrightarrow \begin{cases} -2\sum\limits_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i) = 0 \\ -2\sum\limits_{i=1}^{n} x_i\,(Y_i - \theta_0 - \theta_1 x_i) = 0 \end{cases}.$$

Решая эту систему (а это – линейная система из двух уравнений с двумя неизвестными θ_0 и θ_1), находим, что

$$\theta_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})(Y_i - \overline{Y})}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}, \quad \theta_0 = \overline{Y} - \theta_1\overline{X_1}.$$

Конечно, назвать найденные значения θ_0 и θ_1 оценками можно лишь после того, как мы и правда убедимся, что полученная точка – точка минимума. Это можно сделать, используя какое-нибудь достаточное условие экстремума функции двух переменных, а можно и ограничиться следующим соображением: функция $\varepsilon(\theta_0, \theta_1)$ выпукла вниз, а значит найденная точка, подозрительная на экстремум, и правда является точкой минимума. □

Теперь можно написать, что

$$\widehat{\theta_1} = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})(Y_i - \overline{Y})}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}, \quad \widehat{\theta_0} = \overline{Y} - \theta_1\overline{X_1},$$

где

$$\overline{X_1} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Итак, на основе данных по мобильным телефонам, получаем значения $\widehat{\theta_0} \approx 4.88$, $\widehat{\theta_1} \approx 0.30$. В итоге, функция

$$f(X_1) = 4.88 + 0.30X_1$$

и есть искомая линия линейной регрессии. Построим ее график, он изображен на рисунке 5 синим цветом.

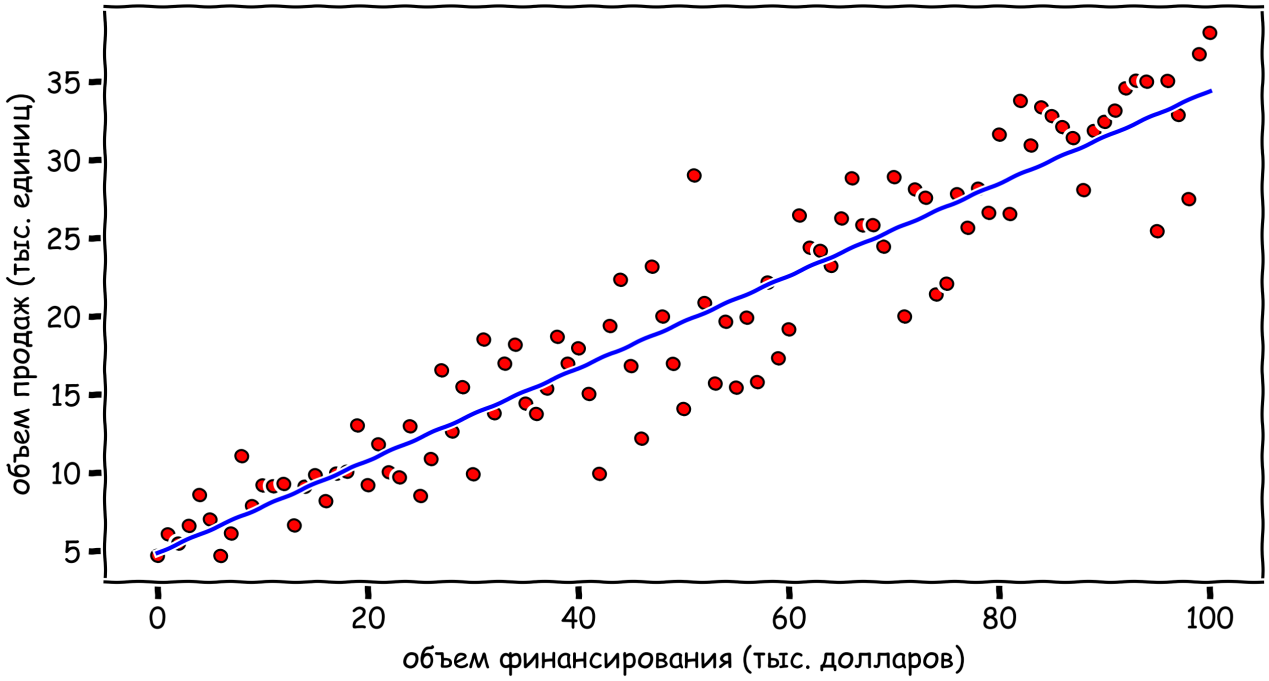

Рис. 5: Зависимость объема продаж мобильных телефонов от затрат на рекламу и регрессия

Как видно из графика, полученная нами модель действительно «неплохо» приближает изображенные данные (ну, строго говоря, что считать критерием плохо или неплохо мы обсудим чуть позже, опять же вернувшись к этим примерам). Кроме того, если провести вертикальные зеленые (параллельные оси Oy) линии от красных точек до синей прямой, то мы получим величины ошибок ε_i (сумму квадратов которых, мы минимизировали), которые показывают отклонение нашей модели от реальных данных, рисунок 6.

На основе данных по продажам самолетов мы получаем значения $\widehat{\theta_0} \approx 5.13$ и $\widehat{\theta_1} \approx 1.34$. Значит, функция

$$f(X_1) = 5.13 + 1.34X_1$$

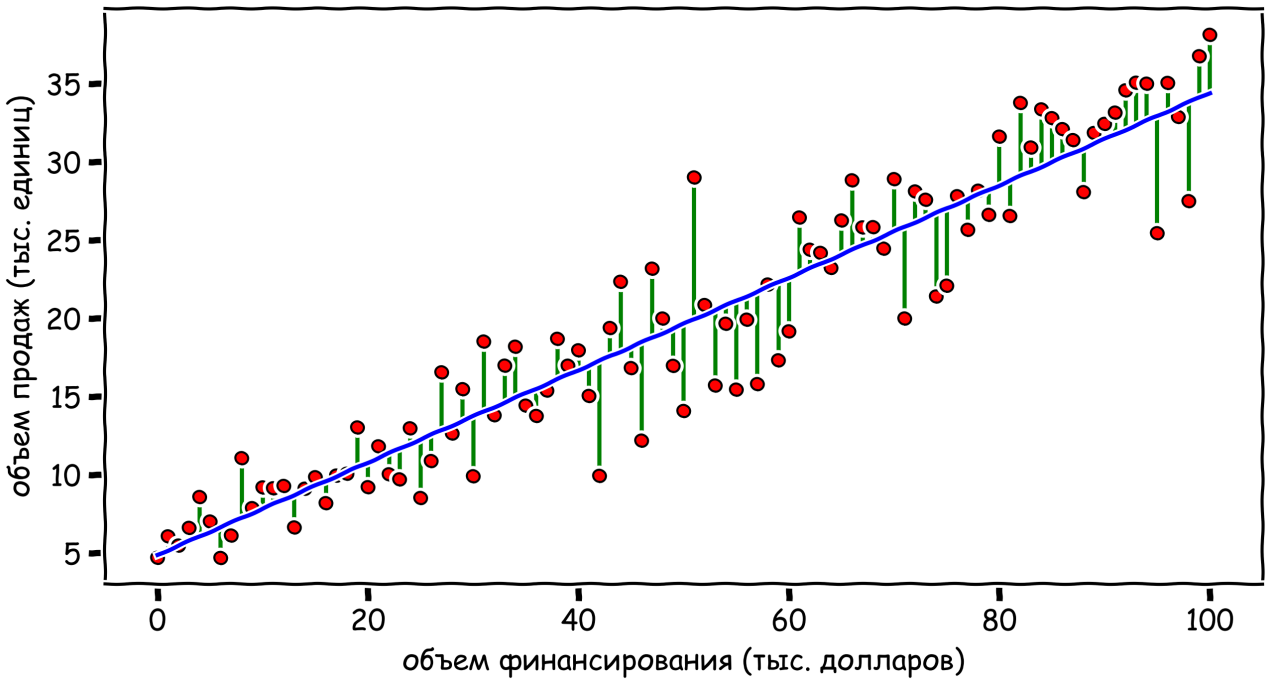

Рис. 6: Зависимость объема продаж мобильных телефонов от затрат на рекламу, регрессия и ошибки

и есть искомая линия линейной регрессии. Построим ее график, он изображен на рисунке 7 синим цветом. Детальное обсуждение точности данной модели, как и модели, полученной ранее, проведем чуть позже.

В этого пункта, еще раз подчеркнем, что, найдя оценки $\widehat{\theta_0}$ и $\widehat{\theta_1}$, предсказание ищется, используя уравнение регрессии

$$Y = \widehat{\theta_0} + \widehat{\theta_1} X_1.$$

Например модель, построенная для мобильных телефонов, с уравнением

$$Y = 4.88 + 0.30 X_1$$

при затратах на рекламу в 150 тысяч долларов дает предсказание по объему продаж в

$$Y = 4.88 + 0.30 \cdot 150 = 49.88$$

тысяч единиц.

1.3 Пример: затраты времени на покупки

Чтобы формулы не казались пугающими, а все увиденное не было чересчур абстрактным, покажем расчеты на конкретном не объемном примере. Пусть, например, имеются данные о том, сколько минут человек находится в продуктовом супермаркете в зависимости от количества приобретаемых им товаров. Данные представим в виде таблицы.

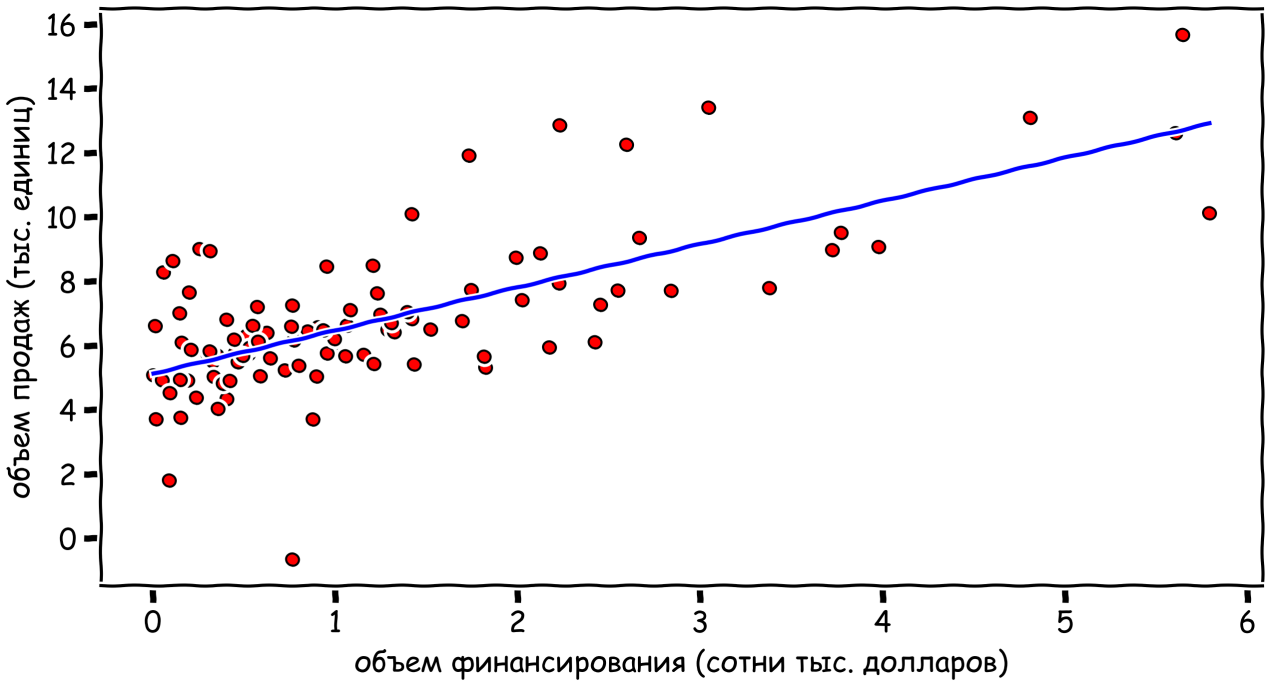

Рис. 7: Зависимость объема продаж самолетов от затрат на рекламу и регрессия

№ наблюдения	Количество выбранных товаров	Время в магазине (мин.)
1	10	15
2	5	12
3	12	18
4	25	30
5	1	3
6	18	20
7	11	14
8	7	10
9	19	20
10	15	13

Предположим, что вам нужно сделать некоторое количество покупок, но вы ограничены во времени. Можно ли спрогнозировать, сколько вам понадобится времени для совершения того или иного количества покупок, опираясь на данные предыдущих походов в магазин? Для перехода к моделированию определим, что является известной переменной, а что откликом. В качестве известной переменной X_1 выберем количество товаров, которое купил человек, а в качестве отклика Y – время, проведенное в магазине. Значит, получаем следующий набор (x_1, Y_1), (x_2, Y_2), $\ldots, (x_{10}, Y_{10})$ пар исходных данных (для удобства приведенных в таблице):

№ наблюдения	Количество выбранных товаров	Время в магазине (мин.)
1	$x_1 = 10$	$Y_1 = 15$
2	$x_2 = 5$	$Y_2 = 12$
3	$x_3 = 12$	$Y_3 = 18$
4	$x_4 = 25$	$Y_4 = 30$
5	$x_5 = 1$	$Y_5 = 3$
6	$x_6 = 18$	$Y_6 = 20$
7	$x_7 = 11$	$Y_7 = 14$
8	$x_8 = 7$	$Y_8 = 10$
9	$x_9 = 19$	$Y_9 = 20$
10	$x_{10} = 15$	$Y_{10} = 13$

Для наглядной иллюстрации изобразим эти данные на рисунке 8. По горизонтальной оси отложены иксы, а по вертикальной – игреки.

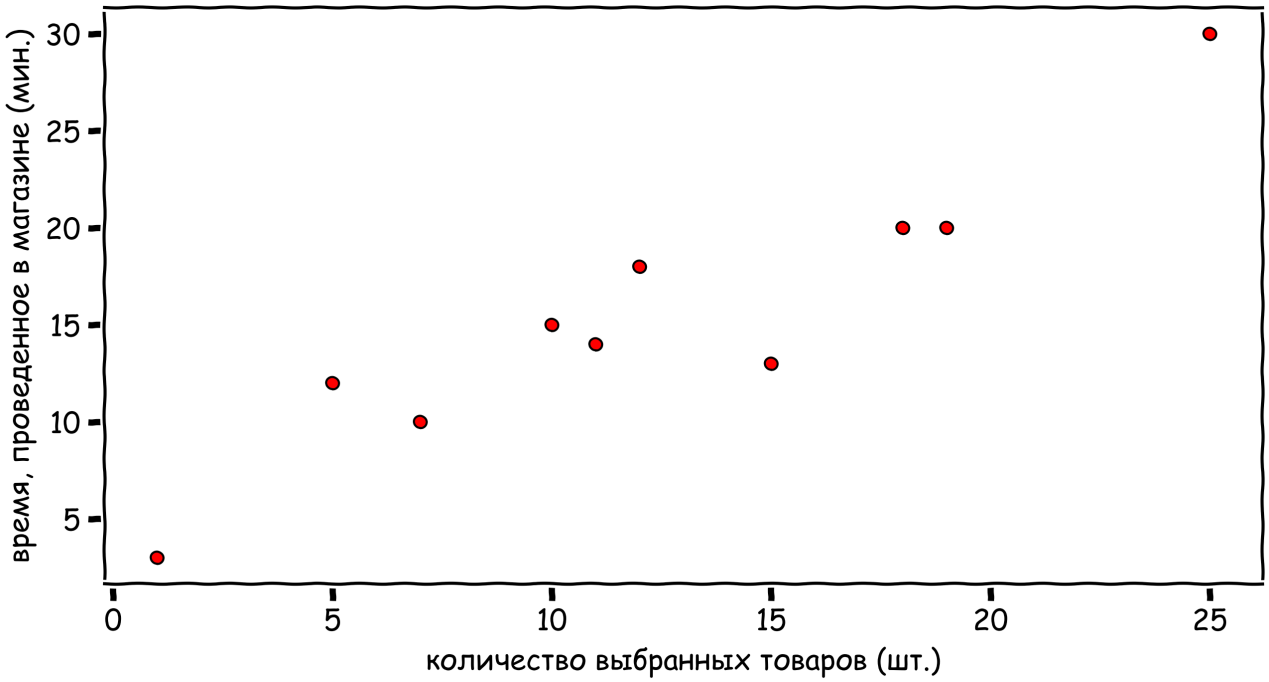

Рис. 8: Зависимость времени, проведенного в магазине, от количества выбранных товаров

Так как в нашем случае $n = 10$, то формулы для $\widehat{\theta_0}$ и $\widehat{\theta_1}$ примут следующий вид:

$$\widehat{\theta_1} = \frac{\sum\limits_{i=1}^{10}(x_i - \overline{X_1})(Y_i - \overline{Y})}{\sum\limits_{i=1}^{10}(x_i - \overline{X_1})^2}, \quad \widehat{\theta_0} = \overline{Y} - \theta_1\overline{X_1},$$

где

$$\overline{X_1} = \frac{1}{10}\sum_{i=1}^{10} x_i, \quad \overline{Y} = \frac{1}{10}\sum_{i=1}^{10} Y_i.$$

Начнем с вычисления последних, итак

$$\overline{X_1} = \frac{1}{10}\,(10 + 5 + 12 + 25 + 1 + 18 + 11 + 7 + 19 + 15) = \frac{123}{10} = 12.3,$$

$$\overline{Y} = \frac{1}{10}\,(15 + 12 + 18 + 30 + 3 + 20 + 14 + 10 + 20 + 13) = \frac{155}{10} = 15.5.$$

Теперь мы можем вычислить $\widehat{\theta_1}$:

$$\widehat{\theta_1} = \frac{(x_1 - \overline{X_1})(Y_1 - \overline{Y}) + (x_2 - \overline{X_1})(Y_2 - \overline{Y}) + ... + (x_{10} - \overline{X_1})(Y_{10} - \overline{Y})}{(x_1 - \overline{X_1})^2 + (x_2 - \overline{X_1})^2 + ... + (x_{10} - \overline{X_1})^2} =$$

$$= \frac{(10 - 12.3)(15 - 15.5) + (5 - 12.3)(12 - 15.5) + ... + (15 - 12.3)(13 - 15.5)}{(10 - 12.3)^2 + (5 - 12.3)^2 + ... + (15 - 12.3)^2} \approx 0.93.$$

Ну а тогда

$$\widehat{\theta_0} \approx 15.5 - 0.93 \cdot 12.3 \approx 4.06.$$

В реальных подсчетах значения лучше не округлять, и подставлять для расчета $\widehat{\theta_0}$ значение $\widehat{\theta_1}$ с как можно большим числом знаков после запятой. Мы округлили $\widehat{\theta_1}$ и нашли приближенное значение $\widehat{\theta_0}$ для наглядности.

Итак, уравнение линейной регрессии имеет следующий вид:

$$Y = 4.06 + 0.93X_1.$$

Построим получившуюся прямую, результат можно видеть на рисунке 9. Вернемся к задаче предсказания. Ответим на вопрос: сколько времени займет поход в магазин, если мы хотим приобрести, например, 27 товаров? Для прогноза достаточно вычислить значение функции $Y = 4.06 + 0.93X_1$ при $X_1 = 27$, то есть

$$4.06 + 0.93 \cdot 27 = 29.17,$$

а значит потребуется чуть больше, чем 29 минут. Иллюстрация прогноза приведена на рисунке 10.

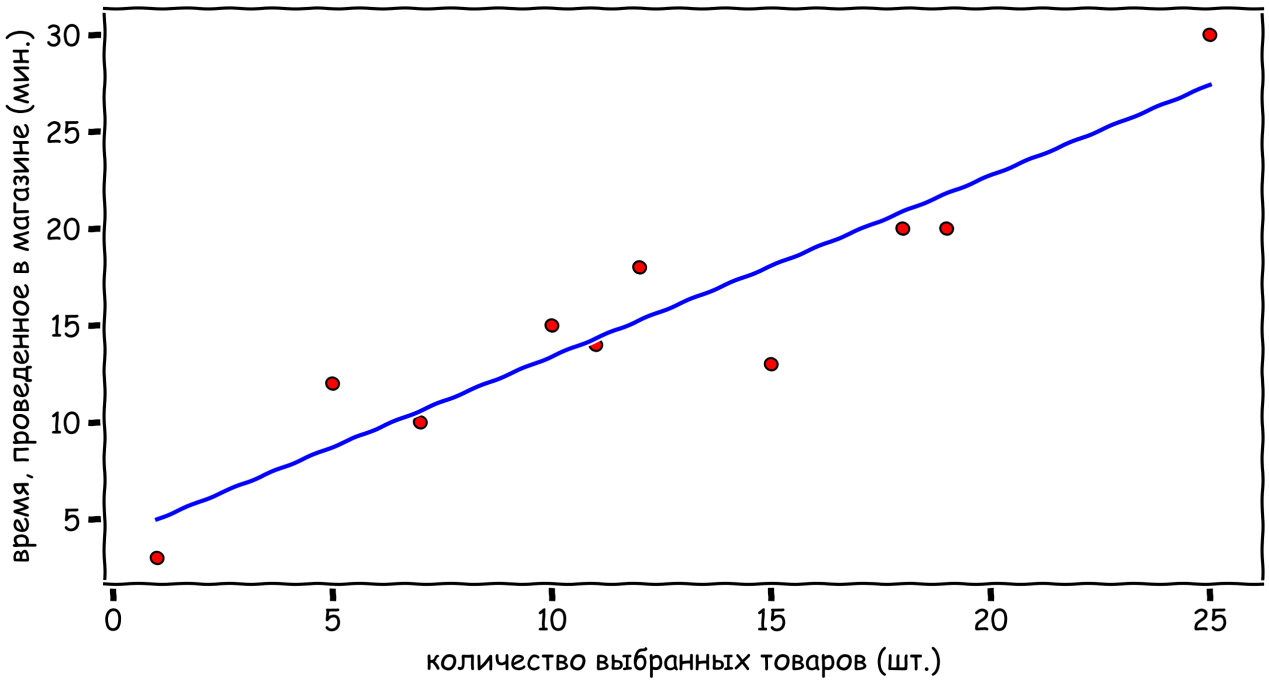

Рис. 9: Зависимость времени, проведенного в магазине, от количества выбранных товаров и регрессия

2 Некоторые статистические характеристики параметров простейшей линейной регрессии

В предыдущем пункте мы решили задачу построения регрессии больше на эвристическом уровне, ведь мы не объяснили, во-первых, почему описанная схема и правда дает хорошую модель, и, в частности, почему применяется метод наименьших квадратов. Во-вторых, за кадром остались рамки применимости модели, в частности вопрос: а что это за случайные ошибки ε, неужели они могут быть любыми? В этом пункте мы подробно обсудим математические детали построенной модели.

2.1 Параметры θ_0 и θ_1 как случайные величины

Напомним предпосылки задачи. Проведя n экспериментов, в которых величина X_1 приняла значения $x_1, x_2, ..., x_n$ (среди которых хотя бы два различны), мы наблюдаем n значений $Y_1, Y_2, ..., Y_n$ нашей случайной величины Y. Так как в результате измерений в эксперименте возникали случайные ошибки, то мы предполагаем, что при каждом i справедливо соотношение

$$Y_i = \theta_0 + \theta_1 x_i + \varepsilon_i$$

при одних и тех же параметрах θ_0 и θ_1.

В этом разделе мы будем исходить из следующих важных предположе-

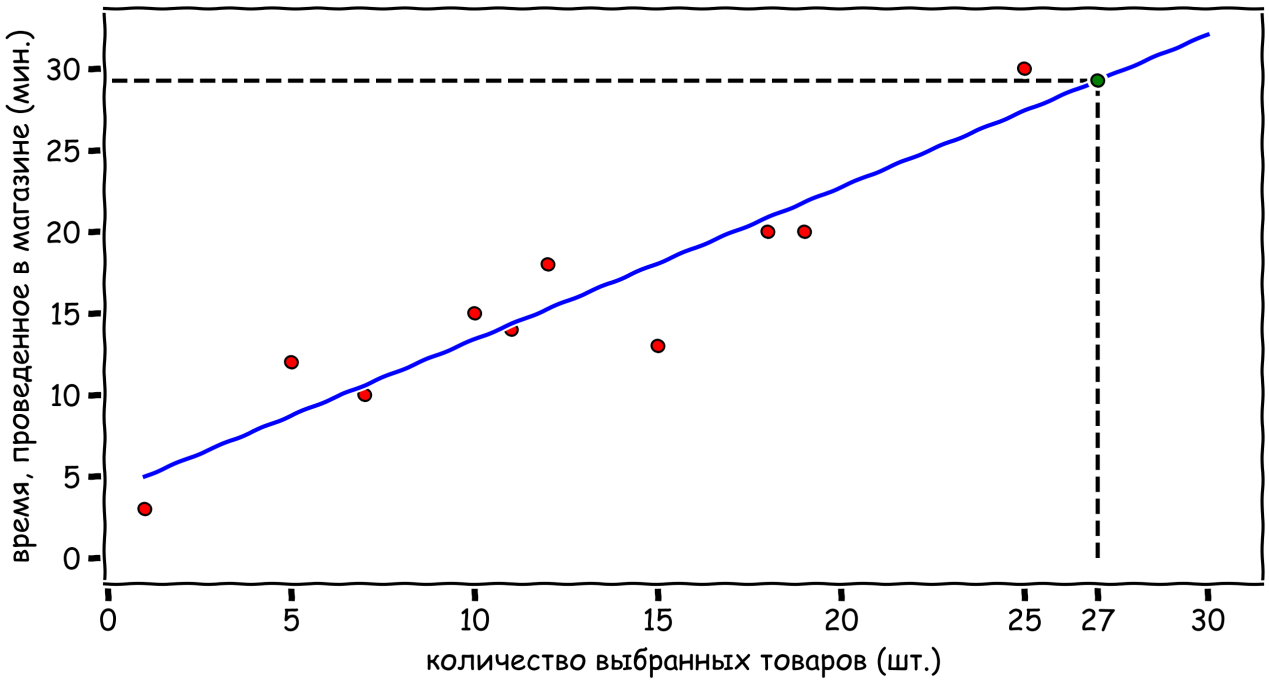

Рис. 10: Зависимость времени, проведенного в магазине, от количества выбранных товаров, регрессия и предсказание

ний (первые три из которых часто называют условиями Гаусса-Маркова):

1. Случайные величины (ошибки) $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ независимы и одинаково распределены;

2. Ошибки не носят систематического характера, то есть $\mathsf{E}\varepsilon_i = 0$, $i \in \{1, 2, ..., n\}$;

3. Дисперсии рошибок одинаковы, то есть $\mathsf{D}\varepsilon_i = \sigma^2 > 0$, $i \in \{1, 2, ..., n\}$ (гомоскедастичность);

4. $\varepsilon_i \sim \mathsf{N}_{0,\sigma^2}$.

Отметим, что даже при таких «жестких» предположениях о распределении случайных ошибок, набор значений $Y_1, Y_2, ..., Y_n$ случайной величины Y не является выборкой в принятом ранее в статистике смысле. Ведь случайные величины Y_i не являются одинаково распределенными, так как, например, их математические ожидания различны, ведь

$$\mathsf{E}Y_i = \mathsf{E}\left(\theta_0 + \theta_1 x_i + \varepsilon_i\right) = \theta_0 + \theta_1 x_i,$$

где последние значения, вообще говоря, не одинаковы.

Теперь мы готовы изучить основные свойства оценок, полученных методом наименьших квадратов. Напомним аналитические выражения для них:

$$\widehat{\theta_1} = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})(Y_i - \overline{Y})}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}, \quad \widehat{\theta_0} = \overline{Y} - \theta_1\overline{X_1},$$

$$\overline{X_1} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Для начала оказывается, что оценки, полученные методом наименьших квадратов, в случае выполнения написанных выше четырех условий совпадают с оценками метода максимального правдоподобия.

Теорема 2.1.1 (О совпадении МНК и ММП оценок) *В предположениях 1-4, сформулированных выше, оценки $\widehat{\theta_0}$ и $\widehat{\theta_1}$ являются оценками максимального правдоподобия параметров θ_0 и θ_1.*

Доказательство. Применим метод максимального правдоподобия. Несмотря на то, что $Y_1, Y_2, ..., Y_n$ – не выборка в привычном смысле, в силу независимости Y_i сам метод, как и его реализация, остаются прежними. Так как

$$Y_i \sim \mathsf{N}_{\theta_0+\theta_1 x_i, \sigma^2},$$

то функция правдоподобия имеет следующий вид

$$f_\theta(\vec{Y}) = f_\theta(Y_1, Y_2, ..., Y_n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum\limits_{i=1}^{n}(Y_i-\theta_0-\theta_1 x_i)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-n/2} e^{-\frac{\varepsilon(\theta_0,\theta_1)}{2\sigma^2}},$$

где

$$\varepsilon(\theta_0, \theta_1) = \sum_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i)^2.$$

Ясно, что максимум функции правдоподобия достигается в случае, когда достигается минимум функции $\varepsilon(\theta_0, \theta_1)$, то есть мы приходим к задаче

$$\arg\min_{\theta_0,\theta_1} \varepsilon(\theta_0, \theta_1) = \arg\min_{\theta_0,\theta_1} \sum_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i)^2,$$

что вторит решаемой задаче в методе наименьших квадратов. □

Замечание 2.1.1 *Заметим, что параметр σ^2 обычно является неизвестным. Так как выборка $Y_1, Y_2, ..., Y_n$ – не выборка в классическом смысле, то в качестве его оценки использовать оценку $S^2(Y)$ или $S_0^2(Y)$, вообще говоря, нельзя.*

Следствие 2.1.2 *Оценкой метода максимального правдоподобия неизвестного параметра σ^2 является следующая случайная величина:*

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2.$$

Доказательство. Как уже было получено ранее,

$$f_{\theta,\sigma^2}(\vec{Y}) = f_{\theta,\sigma^2}(Y_1, Y_2, ..., Y_n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum\limits_{i=1}^{n}(Y_i-\theta_0-\theta_1 x_i)^2}{2\sigma^2}}.$$

Тогда логарифмическая функция правдоподобия перепишется, как

$$L_{\theta,\sigma^2}(\vec{Y}) = -\frac{n}{2}\left(\ln 2\pi + \ln\sigma^2\right) - \frac{\sum\limits_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2}.$$

$$\frac{\partial L_{\theta,\sigma^2}(\vec{Y})}{\partial\sigma^2} = -\frac{n}{2}\cdot\frac{1}{\sigma^2} + \frac{\sum\limits_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^4}.$$

Приравняв последнее выражение к нулю и решив полученное уравнение, получим, что

$$\sigma^2 = \frac{\sum\limits_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i)^2}{n}.$$

С помощью достаточного условия экстремума можно установить, что полученная точка является точкой максимума, а значит

$$\widehat{\sigma^2} = \frac{\sum\limits_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_i)^2}{n}.$$

□

Замечание 2.1.2 *Ясно, что в случае, когда θ_0 и θ_1 неизвестны, оценка переписывается, как*

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1} x_i)^2$$

Теперь рассмотрим свойства оценок $\widehat{\theta_0}$ и $\widehat{\theta_1}$. Полученные ниже факты помогут нам в следующем пункте при построении доверительных интервалов и при проверке гипотез относительно параметров модели. Начнем с $\widehat{\theta_0}$.

Лемма 2.1.1 (О свойствах оценки $\widehat{\theta}_1$) *В предположениях условий Гаусса-Маркова, оценка $\widehat{\theta}_1$ обладает следующими свойствами:*

1. *Она является несмещенной оценкой параметра θ_1;*
2. *Она является эффективной в классе линейных несмещенных оценок;*
3. *Ее дисперсия равна*
$$\mathsf{D}\widehat{\theta}_1 = \frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}.$$
4. *В предположении, что справедливо и четвертое условие, то есть что $\varepsilon_i \sim \mathsf{N}_{0,\sigma^2}$, она имеет нормальное распределение с параметрами θ_1 и $\mathsf{D}\widehat{\theta}_1$, то есть*
$$\widehat{\theta}_1 \sim \mathsf{N}_{\theta_1, \mathsf{D}\widehat{\theta}_1}.$$

Отдельно поясним второе свойство. Оно означает, что ошибка MSE в классе несмещенных линейных оценок минимальна именно на оценках, полученных с помощью метода наименьших квадратов. Так как

$$\mathsf{MSE} = \mathsf{E}||\theta - \widehat{\theta}||^2 = \mathsf{D}\widehat{\theta} + (\mathsf{E}\widehat{\theta} - \theta)^2,$$

а последнее слагаемое для $\widehat{\theta}_1$, в силу несмещенности, равно нулю, то минимальность MSE – суть минимальность дисперсии $\mathsf{D}\widehat{\theta}$ оценки $\widehat{\theta}$. Минимальность дисперсии обеспечивает и минимальность «разброса» оценок от истинного значения параметра.

Доказательство. 1. Пользуясь свойством линейности математического ожидания, получим

$$\mathsf{E}\widehat{\theta}_1 = \frac{1}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2} \sum_{i=1}^{n}(x_i - \overline{X_1})(\mathsf{E}Y_i - \mathsf{E}\overline{Y}).$$

Так как $\mathsf{E}\varepsilon_i = 0$, то

$$\mathsf{E}Y_i = \mathsf{E}\left(\theta_0 + \theta_1 x_i + \varepsilon_i\right) = \theta_0 + \theta_1 x_i,$$

а значит

$$\mathsf{E}\overline{Y} = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}Y_i = \frac{1}{n}\sum_{i=1}^{n}\left(\theta_0 + \theta_1 x_i\right) = \theta_0 + \theta_1\overline{X_1}.$$

Тогда

$$\mathsf{E}Y_i - \mathsf{E}\overline{Y} = \theta_1(x_i - \overline{X_1})$$

и, подставляя это в выражение для $\mathsf{E}\widehat{\theta}_1$, получим

$$\mathsf{E}\widehat{\theta}_1 = \frac{1}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}\sum_{i=1}^{n}\theta_1(x_i - \overline{X_1})^2 = \theta_1.$$

2. Этот пункт в общем случае мы доказывать не будем. Его частный случай следует из свойств оценок метода максимального правдоподобия и предыдущей теоремы.

3. Так как выражение для $\widehat{\theta}_1$ может быть переписано, как

$$\widehat{\theta}_1 = \frac{1}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}\left(\sum_{i=1}^{n}(x_i - \overline{X_1})Y_i - \sum_{i=1}^{n}(x_i - \overline{X_1})\overline{Y}\right),$$

и так как

$$\sum_{i=1}^{n}(x_i - \overline{X_1})\overline{Y} = \overline{Y}\left(\sum_{i=1}^{n}x_i - n\overline{X}\right) = 0,$$

то достаточно исследовать

$$\widehat{\theta}_1 = \frac{1}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}\sum_{i=1}^{n}(x_i - \overline{X_1})Y_i.$$

Пользуясь независимостью ошибок и свойствами дисперсии, получим

$$\mathsf{D}\widehat{\theta}_1 = \frac{1}{\left(\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2\right)^2}\sum_{i=1}^{n}(x_i - \overline{X_1})^2\mathsf{D}Y_i.$$

Так как

$$\mathsf{D}Y_i = \mathsf{D}\left(\theta_0 + \theta_1 x_i + \varepsilon_i\right) = \mathsf{D}\varepsilon_i = \sigma^2,$$

то

$$\mathsf{D}\widehat{\theta}_1 = \frac{\sigma^2}{\left(\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2\right)^2}\sum_{i=1}^{n}(x_i - \overline{X_1})^2 = \frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}.$$

4. Это свойство следует из того, что сумма независимых случайных величин, имеющих нормальное распределение, имеет нормальное распределение, параметры которого вычислены в пунктах 1 и 3.

□

Лемма 2.1.2 (О свойствах оценки $\widehat{\theta}_0$) *В предположениях условий Гаусса-Маркова, оценка $\widehat{\theta}_0$ обладает следующими свойствами:*

1. *Она является несмещенной оценкой параметра θ_0;*
2. *Она является эффективной в классе линейных несмещенных оценок;*
3. *Ее дисперсия равна*
$$\mathsf{D}\widehat{\theta}_0 = \frac{\sigma^2}{n}.$$
4. *В предположении, что справедливо и четвертое условие, то есть что $\varepsilon_i \sim \mathsf{N}_{0,\sigma^2}$, она имеет нормальное распределение с параметрами θ_0 и $\mathsf{D}\widehat{\theta}_0$, то есть*
$$\widehat{\theta}_0 \sim \mathsf{N}_{\theta_0,\mathsf{D}\widehat{\theta}_0}.$$
5. *В предположении, что справедливо и четвертое условие, то есть что $\varepsilon_i \sim \mathsf{N}_{0,\sigma^2}$, она является оценкой максимального правдоподобия параметра θ_0.*

Доказательство. 1. Из свойства линейнойсти математического ожидания, получим

$$\mathsf{E}\widehat{\theta_0} = \mathsf{E}\left(\overline{Y} - \theta_1\overline{X_1}\right) = \mathsf{E}\overline{Y} - \theta_1\overline{X_1} = \frac{1}{n}\sum_{i=1}^{n}(\theta_0 + \theta_1 x_i) - \theta_1\overline{X_1} =$$

$$= \theta_0 + \theta_1\overline{X_1} - \theta_1\overline{X_1} = \theta_0.$$

2. Этот пункт в общем случае мы доказывать не будем. Его частный случай следует из свойств оценок метода максимального правдоподобия и теоремы о связи оценок МНК и ММП.
3. Используя свойства дисперсии, получим

$$\mathsf{D}\widehat{\theta_0} = \mathsf{D}(\overline{Y} - \theta_1\overline{X_1}) = \mathsf{D}\overline{Y} = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{D}Y_i.$$

Так как

$$\mathsf{D}Y_i = \mathsf{D}(\theta_0 + \theta_1 x_i + \varepsilon_i) = \mathsf{D}\varepsilon_i = \sigma^2,$$

то

$$\mathsf{D}\widehat{\theta_0} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

4. Это свойство следует из того, что сумма независимых нормально распределенных случайных величин имеет нормальное распределение, и из вычислений в пунктах 1 и 3. □

Замечание 2.1.3 *В последней теореме вывод дисперсии оценки $\widehat{\theta_0}$ существенным образом опирается на то, что θ_1 – известное число. Если же θ_1 оценивается при помощи $\widehat{\theta_1}$, то вычисления становятся немного сложнее и приводят к следующему выражению для дисперсии*

$$\mathsf{D}\widehat{\theta_0} = \mathsf{D}\overline{Y} + \mathsf{D}\widehat{\theta_1} \cdot \overline{X_1}^2,$$

откуда

$$\mathsf{D}\widehat{\theta_0} = \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}\overline{X_1}^2 = \sigma^2\left(\frac{1}{n} + \frac{\overline{X_1}^2}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}\right)$$

2.2 Построение доверительных интервалов для коэффициентов регрессии

Итак, перед тем как сформулировать основную теорему о доверительных интервалах, сначала вернемся к оценке неизвестного параметра σ^2. Напомним, что согласно ММП, она имеет вид

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2.$$

Замечание 2.2.1 *Можно показать, что в предположении условий 1-4, обсужденных ранее,*

$$\frac{n}{\sigma^2}\widehat{\sigma^2} \sim \chi^2_{n-2}.$$

По сути дела, 2 степени свободы «забираются» из-за того, что мы нее знаем ни параметр θ_0, ни параметр θ_1, а лишь оцениваем их. Пользуясь этим, согласно свойству линейности математического ожидания и согласно тому, что математическое ожидание случайной величины, имеющей распределение хи-квадрат с $(n-2)$ степенями свободы, равно $(n-2)$, получаем

$$\mathsf{E}\left(\frac{n}{\sigma^2}\widehat{\sigma^2}\right) = \frac{n}{\sigma^2}\mathsf{E}\widehat{\sigma^2} = (n-2).$$

А тогда

$$\mathsf{E}\widehat{\sigma^2} = \frac{n-2}{n}\sigma^2,$$

и полученная нами оценка $\widehat{\sigma^2}$ является смещенной, но асимптотически несмещенной, ведь.

$$\lim_{n\to+\infty}\frac{n-2}{n} = 1.$$

Несмещенная же оценка получается так:

$$\widehat{\sigma_0^2} = \frac{n}{n-2}\widehat{\sigma^2} = \frac{n}{n-2} \cdot \frac{\sum\limits_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2}{n} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2.$$

Так как зачастую истинное значение параметра σ^2 неизвестно, то его приходится оценивать, а значит приходится оценивать и дисперсии оценок $\widehat{\theta_0}$ и $\widehat{\theta_1}$, полученных ранее. Итак, оценим «разброс» среди возможных оценок $\widehat{\theta_0}$ и $\widehat{\theta_1}$.

Определение 2.2.1 *Величины*

$$\mathsf{SE}(\widehat{\theta_0}) = \sqrt{\frac{\sum\limits_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2}{n-2}} \cdot \sqrt{\frac{1}{n} + \frac{\overline{X_1}^2}{\sum\limits_{i=1}^{n}\left(x_i - \overline{X_1}\right)^2}},$$

$$\mathsf{SE}(\widehat{\theta_1}) = \sqrt{\frac{\sum\limits_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2}{n-2}} \cdot \sqrt{\frac{1}{\sum\limits_{i=1}^{n}\left(x_i - \overline{X_1}\right)^2}}$$

называются стандартными ошибками (standard error) оценок $\widehat{\theta_0}$ и $\widehat{\theta_1}$, соответственно.

Замечание 2.2.2 *Полезно понимать, откуда получаются и что обозначают только что введенные стандартные ошибки. Так как*

$$\mathsf{D}\widehat{\theta_0} = \sigma^2\left(\frac{1}{n} + \frac{\overline{X_1}^2}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}\right),$$

и

$$\mathsf{D}\widehat{\theta_1} = \frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2},$$

то введенные стандартные ошибки есть не что иное, как оценки среднеквадратических отклонений $\widehat{\theta_0}$ и $\widehat{\theta_1}$ при замене σ^2 на ее несмещенную оценку, полученную ранее, то есть на оценку

$$\widehat{\sigma_0^2} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2.$$

На основе стандартных ошибок можно построить так называемый доверительный интервал. Напомним определение.

Определение 2.2.2 *Пусть* $0 < \varepsilon < 1$. *Доверительный интервал* $(\theta^-,\ \theta^+)$ *уровня доверия* $1-\varepsilon$ *– это интервал, в который с вероятностью* $1-\varepsilon$ *попадет реальное значение параметра.*

На практике часто рассматривают значения $\varepsilon = 0.1$, $\varepsilon = 0.05$ или $\varepsilon = 0.01$.

Теорема 2.2.1 (Доверительные интервалы для θ_0 и θ_1) *В предположении условий 1 - 4, доверительный интервал уровня доверия* $(1-\varepsilon)$ *для параметра* θ_0 *– это интервал*

$$\left(\widehat{\theta_0} - t_{1-\varepsilon/2}\cdot \mathsf{SE}(\widehat{\theta_0}),\ \widehat{\theta_0} + t_{1-\varepsilon/2}\cdot \mathsf{SE}(\widehat{\theta_0})\right),$$

где $t_{1-\varepsilon/2}$ *– это* $(1-\varepsilon/2)$ *квантиль распределения Стьюдента с* $(n-2)$ *степенями свободы.*

Аналогично, доверительный интервал уровня доверия $(1-\varepsilon)$ *для параметра* θ_1 *– это интервал*

$$\left(\widehat{\theta_1} - t_{1-\varepsilon/2}\cdot \mathsf{SE}(\widehat{\theta_1}),\ \widehat{\theta_1} + t_{1-\varepsilon/2}\cdot \mathsf{SE}(\widehat{\theta_1})\right),$$

где $t_{1-\varepsilon/2}$ *–* $(1-\varepsilon/2)$ *квантиль распределения Стьюдента с* $(n-2)$ *степенями свободы.*

Доказательство. Докажем, например, второе соотношение. Первое получается аналогичным образом. Как мы знаем (из леммы о свойствах оценки $\widehat{\theta_1}$),

$$\widehat{\theta_1} \sim \mathsf{N}_{\theta_1,\ \frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i-\overline{X_1})^2}}.$$

Тогда, используя свойства линейных преобразований,

$$\frac{\theta_1 - \widehat{\theta_1}}{\sqrt{\frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i-\overline{X_1})^2}}} \sim \mathsf{N}_{0,1}.$$

Кроме того, согласно замечанию в начале данного пункта,

$$\frac{n\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-2}.$$

Тогда, в силу независимости, случайная величина

$$t_{n-2} = \left(\frac{\theta_1 - \widehat{\theta_1}}{\sqrt{\frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \overline{X_1})^2}}} \right) : \left(\sqrt{\frac{n\widehat{\sigma^2}}{\sigma^2(n-2)}} \right) \sim \mathsf{T}_{n-2}$$

имеет распределение Стьюдента с $(n-2)$-мя степенями свободы. Последняя же эквивалентным образом переписывается в виде

$$t_{n-2} = \frac{\theta_1 - \widehat{\theta_1}}{\mathsf{SE}(\widehat{\theta_1})}.$$

Дальнейшее построение доверительного интервала стандартно. □

2.3 Немного об интерпретации доверительных интервалов

Возвратимся теперь к примерам, касающимся продаж телефонов и самолетов, с которых мы начали данную лекцию.

В примере с мобильными телефонами, доверительный интервал $\left(\theta_0^-,\ \theta_0^+\right)$ для θ_0 при $\varepsilon = 0.1$ имеет вид

$$\left(\theta_0^-,\ \theta_0^+\right) = (3.88,\ 5.87)\,,$$

а для θ_1 имеет вид

$$\left(\theta_1^-,\ \theta_1^+\right) = (0.28,\ 0.31)\,.$$

Подробный пример расчета мы увидим чуть позже, а сейчас давайте вдумаемся в смысл полученных интервалов. Напомним для наглядности, что уравнение регрессии таково:

$$Y = \theta_0 + \theta_1 X_1.$$

Итак, исходя из расчетов, можно сделать вывод, что при отсутствии расходов на рекламу (то есть при $X_1 = 0$) продажи, в среднем, упадут до $3.88 - 5.87$ тысяч единиц (так как прогнозируемое значение $Y = \theta_0$). При этом, за каждую потраченную на рекламу тысячу долларов объем продаж в среднем увеличится на $0.28 - 0.31$ тысяч единиц.

В примере с самолетами, доверительный интервал уровня доверия 0.9 для θ_0 имеет вид

$$\left(\theta_0^-,\ \theta_0^+\right) = (4.73,\ 5.52)$$

а для θ_1 имеет вид

$$\left(\theta_1^-,\ \theta_1^+\right) = (1.12,\ 1.57).$$

Проанализировав эти результаты, можно сделать вывод, что при отсутствии рекламы продажи, в среднем, упадут до $4.73 - 5.52$ тысяч единиц. При этом за каждую потраченную на рекламу сотню тысяч долларов объем продаж в среднем увеличится на $1.12 - 1.57$ тысячу единиц.

2.4 Доверительные интервалы для примера

Вернемся к нашему примеру со временем, проведенном в магазине, и вычислим доверительные интервалы для параметров модели. Запишем формулы в случае десяти исходных данных:

$$\mathsf{SE}(\widehat{\theta_0}) = \sqrt{\frac{\sum\limits_{i=1}^{10}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2}{10-2}} \cdot \sqrt{\frac{1}{10} + \frac{\overline{X_1}^2}{\sum\limits_{i=1}^{10}\left(x_i - \overline{X_1}\right)^2}},$$

$$\mathsf{SE}(\widehat{\theta_1}) = \sqrt{\frac{\sum\limits_{i=1}^{10}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2}{10-2}} \cdot \sqrt{\frac{1}{\sum\limits_{i=1}^{10}\left(x_i - \overline{X_1}\right)^2}},$$

$$\overline{X_1} = \frac{1}{10}\sum_{i=1}^{10} x_i.$$

Опишем, как находить $t_{1-\varepsilon/2}$ из формулы для доверительного интервала. Итак, у нас $n = 10$, то есть десять степеней свободы. Пусть $\varepsilon = 0.1$, тогда $1 - \varepsilon/2 = 0.95$, $n - 2 = 8$, значит в таблице, которую можно найти в дополнительных материалах, находим значение на пересечении восьмой строки и столбца, соответствующего вероятности 0.95. В нашем случае получаем $t_{0.95} \approx 1.86$. Так как $\mathsf{SE}(\widehat{\theta_0}) = 0.91$ и $\mathsf{SE}(\widehat{\theta_1}) = 0.13$, то

$$\left(\theta_0^-,\ \theta_0^+\right) = (2.37,\ 5.75)$$

и

$$\left(\theta_1^-,\ \theta_1^+\right) = (0.69,\ 1.17)\,.$$

Какой же можно сделать вывод из проделанных расчетов? Рассмотрим сначала второй интервал. В среднем, увеличение количества товаров на один, в 90% случаев (ведь мы взяли $\varepsilon = 0.1$, а значит вероятность $1 - \varepsilon = 0.9$)

увеличивает время пребывания в магазине от 0.69 минут до 1.17 минут. Первый же интервал показывает, что, проведя в магазине ноль минут, можно в среднем купить $3-4$ товара. Такая аномалия обусловлена как ошибкой в модели (зависимость не абсолютно линейная), так и маленьким количеством реальных данных. Если предположить, что мы проводим в магазине хотя бы одну минуту, то аномалия пропадает.

2.5 Проверка гипотез

Стандартные ошибки также используются в так называемой задаче проверки гипотез. Одна из самых часто проверяемых гипотез – так называемая гипотеза статистической значимости параметра $\widehat{\theta_1}$, формулируется следующим образом:

$$\mathsf{H}_0: \text{ Между } X_1 \text{ и } Y \text{ нет зависимости.}$$

Альтернативная ей гипотеза такова

$$\mathsf{H}_a: \text{ Между } X_1 \text{ и } Y \text{ есть зависимость.}$$

С точки зрения математики, нулевая и альтернативная гипотезы говорят не что иное, как

$$\mathsf{H}_0: \ \theta_1 = 0,$$
$$\mathsf{H}_a: \ \theta_1 \neq 0.$$

Действительно, в случае $\theta_1 = 0$ модель переписывается в виде $Y = \theta_0$ и значения X не учитываются вовсе.

Для проверки гипотезы необходимо определить, насколько значение нашей оценки $\widehat{\theta_1}$ истинного параметра θ_1 далеко от нуля. Ясно, что это зависит от стандартной ошибки $\mathsf{SE}(\widehat{\theta_1})$. Если последняя мала, то даже достаточно малые значения θ_1^* могут доказывать, что $\theta_1 \neq 0$. Если же ошибка велика, то и значение $|\widehat{\theta_1}|$ должно быть велико, чтобы отвергнуть нулевую гипотезу. На практике обычно используют t-критерий Стьюдента. Для этого вычисляют статистику

$$t = \frac{|\widehat{\theta_1} - 0|}{\mathsf{SE}(\widehat{\theta_1})} = \frac{|\widehat{\theta_1}|}{\mathsf{SE}(\widehat{\theta_1})}.$$

Сравнивая фактическое и табличное значение $t_{1-\varepsilon/2}$ на уровне доверия $1-\varepsilon$ с числом степеней свободы $(n-2)$ принимается решение:

1. Если $t_{1-\varepsilon/2} < t$, то гипотеза H_0 отклоняется и оценка $\widehat{\theta_1}$ признается статистически значимой на уровне значимости ε.

2. Если $t_{1-\varepsilon/2} \geq t$, то гипотеза H_0 принимается и оценка $\widehat{\theta_1}$ признается статистически незначимой на уровне значимости ε.

Конечно, нас интересует, чтобы выполнялся первый пункт, иначе наша модель, с точки зрения статистики, не отражает реальной зависимости между переменными.

В случае с мобильными телефонами при $\varepsilon = 0.1$ мы получаем значение $t = 28.49$, что больше значения $t_{1-\varepsilon/2} \approx 1.66$ из таблицы, значит гипотеза H_0 отклоняется и принимается альтернативная гипотеза H_1. В случае с самолетами мы получаем значение $t = 9.81$, что снова больше значения $t_{1-\varepsilon/2}$ из таблицы, значит гипотеза H_0 отклоняется и принимается альтернативная гипотеза H_1.

Замечание 2.5.1 *Конечно, можно проверять и гипотезу о равенстве θ_1 какому-то конкретному значению. Для этого имеет смысл использовать статистику*

$$t = \frac{|\widehat{\theta_1} - \theta_1|}{\mathsf{SE}(\widehat{\theta_1})}.$$

Дальнейшие действия абсолютно такие же, как в приведенном алгоритме.

Замечание 2.5.2 *Аналогичным образом можно проверять гипотезы и относительно значений параметра θ_0. Подробнее об этом мы поговорим в разделе, посвященном множественной регрессии.*

2.6 Проверка гипотез для примера

Все необходимые значения для подсчета уже вычислены. Пусть, опять же, $\varepsilon = 0.1$. В нашем случае

$$t = \frac{|\widehat{\theta_1}|}{\mathsf{SE}(\widehat{\theta_1})} = \frac{0.93}{0.13} \approx 7.15.$$

что больше, чем 1.86, значит гипотеза H_0 отклоняется и принимается альтернативная гипотеза H_1. Тем самым установлен ненулевой отклик на предиктор, зависимость имеется.

2.7 Оценка точности модели

Если нулевая гипотеза отвергнута в пользу альтернативной гипотезы, довольно естественно задаться целью определить степень того, насколько модель подходит под данные. Обычно такую «оценку» линейной регрессии дают две величины: среднеквадратическое отклонение остатков (RSE – residual standard error) и R^2 статистика.

В модели мы четко видим, что каждый опыт наделен некоторой ошибкой ε. Из-за этой ошибки, даже зная реальные значения коэффициентов θ_0 и θ_1, мы не сможем точно предсказать значение Y, зная значение X_1. RSE –

оценка среднего квадратичного отклонения σ^2 ошибки ε. Грубо говоря, она показывает насколько построенная модель отличается от «настоящей» модели и оценивает «усредненный» корень из суммы квадратов ошибок модели. RSE, как мы уже знаем, может быть вычислена по формуле

$$\widehat{\sigma_0} = \mathsf{RSE} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2}.$$

Так как RSE измеряется в тех же единицах, что и Y, не всегда понятно, хороший ли получается показатель. R^2 статистика, в отличие от RSE, величина безразмерная и лежит между нулем и единицей. Для вычисления R^2, используют формулу

$$\mathsf{R}^2 = 1 - \frac{\sum\limits_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_i)^2}{\sum\limits_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}.$$

Ясно, что если построенная идеально соответствует исходным данным, то все слагаемые в последней дроби числителя равны нулю и, тогда и только тогда, $\mathsf{R}^2 = 1$ – модель идеальна. Наоборот, если $\widehat{\theta_1} = 0$, то есть модель не зависит от X_1, то $\mathsf{R}^2 = 0$ (так как в случае $\widehat{\theta_1} = 0$ имеем $\widehat{\theta_0} = \overline{Y}$), и модель совершенно несостоятельна, так как не отражает никакой зависимости между откликом и предиктором.

Несмотря на то, что значения R^2 статистики лежат между нулем и единицей, мы все равно не можем сказать, какое значение R^2 является хорошим. Например, в некоторых задачах физики мы точно знаем, что зависимость линейна с незначительной ошибкой, и будем ожидать коэффициент очень близким к единице, а маленькое значение коэффициента будет свидетельствовать о серьезной проблеме в эксперименте, из которого брались данные. Во многих же других областях, как биология, маркетинг и проч., линейная модель является довольно грубой аппроксимацией данных, и ошибки часто велики.

Напомним, что выборочная корреляция X_1 и Y определяется, как

$$r\left(X_1, Y\right) = \frac{\sum\limits_{i=1}^{n}\left(x_i - \overline{X_1}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum\limits_{i=1}^{n}\left(x_i - \overline{X_1}\right)^2}\sqrt{\sum\limits_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}}$$

Можно показать, что, как оказывается, $\mathsf{R}^2 = r^2(Y, \widehat{\theta_0} + \widehat{\theta_1}X_1)$.

В случае примера с мобильными телефонами мы получаем $\mathsf{RSE} = 3.04$ и $\mathsf{R}^2 = 0.89$, из чего, согласно вышесказанному, мы можем сделать вывод, что с точки зрения статистики наша модель работает неплохо и действительно может описывать рассматриваемую зависимость.

В случае примера с самолетами мы получаем $\mathsf{RSE} = 1.73$ и $\mathsf{R}^2 = 0.49$. Характеристики данной модели намного хуже, чем предыдущей.

2.8 Оценка точности модели для примера

Для нашего примера формулы переписываются в следующем виде:

$$\mathsf{RSE} = \sqrt{\frac{1}{10-2}\sum_{i=1}^{10}(Y_i - 4.06 - 0.93 \cdot x_i)^2},$$

$$\mathsf{R}^2 = 1 - \frac{\sum\limits_{i=1}^{10}\left(Y_i - 4.06 - 0.93 \cdot x_i\right)^2}{\sum\limits_{i=1}^{10}\left(Y_i - \overline{Y}\right)^2}.$$

Проведя вычисления, получаем

$$\mathsf{RSE} = 2.77, \quad \mathsf{R}^2 = 0.87,$$

что свидетельствует о том, что модель, с точки зрения статистики, хорошая.

3 Множественная линейная регрессия

Простейшая линейная регрессия позволяет решить задачу предсказания значения одной переменной, зная другую. Но на практике интересующее нас значение часто зависит более, чем от одной переменной. Скажем объемы продаж компании зависят как от того, сколько потрачено на рекламу телефонов, так и от того, сколько потрачено на рекламу самолетов. Как нам расширить наш анализ на большее количество переменных?

3.1 Основные определения и матричные обозначения

Достаточно подробно изучив одномерную регрессию, по аналогии мы можем записать модель множественной линейной регрессии в следующем виде

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_p X_p + \varepsilon,$$

где $X_1, X_2, ..., X_p$ – (неслучайные) входные данные, по которым мы пытаемся определить (случайную) переменную Y (отклик), а ε – некоторая случайная ошибка. Итак, зависимость между Y и X_1, X_2, ..., X_p предполагается линейной с точностью до некоторой ошибки ε.

Определение 3.1.1 *Модель, описываемая зависимостью*

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_p X_p + \varepsilon,$$

где $\theta_0, \theta_1, ..., \theta_p$ – числовые параметры, $X_1, X_2, ..., X_p$ – неслучайные параметры, значения которых либо заданы, либо наблюдаются (иначе говоря – известны), ε – случайная ошибка, называется моделью множественной линейной регрессии.

Часто используют и следующие два определения.

Определение 3.1.2 *Функция*

$$f(X_1, X_2, ..., X_p) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_p X_p$$

в модели множественной линейной регрессии называется линией регрессии Y на X_1, X_2, ..., X_p.

Определение 3.1.3 *Уравнение*

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_p X_p$$

в модели множественной линейной регрессии называется уравнением регрессии Y на X_1, X_2, ..., X_p.

Как же определить коэффициенты модели на конкретных наблюдаемых значениях X_1, X_2, ..., X_p и Y? Давайте опишем схему подробнее. Начнем же снова с того, что аккуратно выпишем: а что дано?

Ясно, что на каждом i-ом наблюдении мы получаем значение Y_i по значениям $(x_{i1}, x_{i2}, ..., x_{ip})$ предикторов $X_1, X_2, ..., X_p$, соответственно, а в каждом наблюдении выполняется равенство

$$Y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + ... + \theta_p x_{ip} + \varepsilon_i, \quad i \in \{1, 2, ..., n\}$$

Далее нам будет удобно работать в матричных обозначениях. Введем дополнительные обозначения $x_{10} = x_{20} = ... = x_{n0} = 1$ и

$$X = \begin{pmatrix} x_{10} & x_{11} & \dots & x_{1p} \\ x_{20} & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \dots & x_{np} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \Theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}.$$

Тогда нашу модель в матричном виде можно переписать, как

$$Y = X \cdot \Theta + \Sigma.$$

Замечание 3.1.1 *По сути дела, введение дополнительных обозначений – суть введение еще одного предиктора X_0, который в каждом наблюдении принимает одно и то же значение, равное 1. При таком соглашении ясно, что рассматриваемая нами модель*

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_p X_p + \varepsilon$$

эквивалентно переписывается в виде

$$Y = \theta_0 X_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_p X_p + \varepsilon.$$

Отметим также следующее часто встречающееся определение.

Определение 3.1.4 *Матрица X, введенная выше, часто называется регрессором.*

Теперь перейдем к поиску коэффициентов множественной регрессии.

3.2 МНК для множественной регрессии

Для нахождения оценок неизвестных параметров аналогично тому, что было сделано в одномерном случае, применим МНК. Тем самым оценки $\widehat{\theta_0}, \widehat{\theta_1}, ..., \widehat{\theta_p}$ коэффициентов $\theta_0, \theta_1, ..., \theta_p$ находятся из решения задачи минимизации функции $\varepsilon(\theta_0, \theta_1, ..., \theta_p)$, зависящей уже от $(p+1)$-ой переменной

$$\varepsilon(\theta_0, \theta_1, ..., \theta_p) = \sum_{i=1}^{n} \left(Y_i - \theta_0 - \theta_1 x_{i1} - \theta_2 x_{i2} - ... - \theta_p x_{ip}\right)^2.$$

Так как на самом деле нам нужно не наименьшее значение функции, а коэффициенты, ее минимизирующие, то решается задача

$$\arg\min_{\theta_0,...,\theta_p} \sum_{i=1}^{n} \left(Y_i - \theta_0 - \theta_1 x_{i1} - \theta_2 x_{i2} - ... - \theta_p x_{ip}\right)^2.$$

Определение 3.2.1 *Оценкой метода наименьших квадратов для неизвестных параметров θ_0, θ_1, ...,θ_p модели множественной регрессии называется набор значений параметров, минимизирующий выражение*

$$\varepsilon(\theta_0, \theta_1, ..., \theta_p) = \sum_{i=1}^{n} \left(Y_i - \theta_0 - \theta_1 x_{i1} - \theta_2 x_{i2} - ... - \theta_p x_{ip}\right)^2.$$

Поставленную задачу можно было бы решить ровно так, как мы делали ранее, но вычислений становится больше, и они становятся более громоздкими. Сформулируем окончательную теорему и воспользуемся геометрическими соображениями для ее пояснения.

Теорема 3.2.1 *Пусть столбцы регрессора X линейно независимы, а $n > (p+1)$ (количество наблюдений больше, чем количество неизвестных параметров модели). Минимум функции*

$$\varepsilon(\theta_0, \theta_1, ..., \theta_p) = \sum_{i=1}^{n} \left(Y_i - \theta_0 - \theta_1 x_{i1} - \theta_2 x_{i2} - ... - \theta_p x_{ip}\right)^2$$

единственен и достигается при

$$\Theta = \left(X^T X\right)^{-1} X^T Y.$$

Доказательство. Проведем обоснование для случая $p = 1$, то есть при двух неизвестных параметрах θ_0 и θ_1. В общем случае обоснование такое же, но становится менее геометричным.

Пусть X_0, X_1 – столбцы регрессора X, они линейно независимы, а значит порождают двумерное пространство $L = \mathbb{R}^2$. Кроме того,

$$X\Theta = X_0\theta_0 + X_1\theta_1 \in L.$$

Из геометрических соображений (рисунок **??**) ясно, что, согласно методу наименьших квадратов, мы минимизируем квадрат длины $Y - X\Theta$ (последний вектор обозначен пунктиром). Понятно, что длина (а значит и квадрат длины) минимален, когда вектор $Y - X\Theta$ ортогонален L, то есть ортогонален каждому вектору X_0 и X_1:

$$(Y - X\Theta) \perp X_i,\ i = 0, 1.$$

Иными словами,

$$X^T(Y - X\Theta) = 0 \Leftrightarrow X^TY - X^TX\Theta = 0.$$

Так как $\det(XX^T) \neq 0$, то

$$\Theta = (X^TX)^{-1}X^TY.$$

□

Итого,

$$\widehat{\Theta} = (X^TX)^{-1}X^TY.$$

Замечание 3.2.1 *Конечно, при решении практических задач полученные формулы в явном виде применять приходится редко – соответствующие функции вшиты в большинство математических пакетов.*

Зная оценки коэффициентов модели, предсказание может быть сделано в соответствии с формулой

$$Y = \widehat{\theta_0} + \widehat{\theta_1}X_1 + \widehat{\theta_2}X_2 + ... + \widehat{\theta_p}X_p.$$

Обратимся к примеру построения многомерной регрессии на основе ранее рассмотренных примеров. Рассмотрим вот какой вопрос: как зависит суммарный объем продаж самолетов и телефонов от вклада в рекламу каждого товара по отдельности? Распределение исходных данных можно увидеть на рисунке 11. Ясно, что в нашем случае количество предикторов равно двум, а значит модель имеет следующий вид:

$$Y = \theta_0 + \theta_1X_1 + \theta_2X_2 + \varepsilon.$$

Найдем оценки $\widehat{\theta_0}$, $\widehat{\theta_1}$ и $\widehat{\theta_2}$ неизвестных коэффициентов θ_0, θ_1 и θ_2. Используя вышеприведенные формулы, имеем

$$\widehat{\theta_0} = 27.20, \quad \widehat{\theta_1} = 1.08, \quad \widehat{\theta_2} = 0.86,$$

и предсказание осуществляется из соотношения

$$Y = 27.20 + 1.08 \cdot X_1 + 0.86 \cdot X_2.$$

Попытки изобразить линию регрессии (плоскость) и разброс данных с разных ракурсов приведены на рисунках 12, 13 и 14.

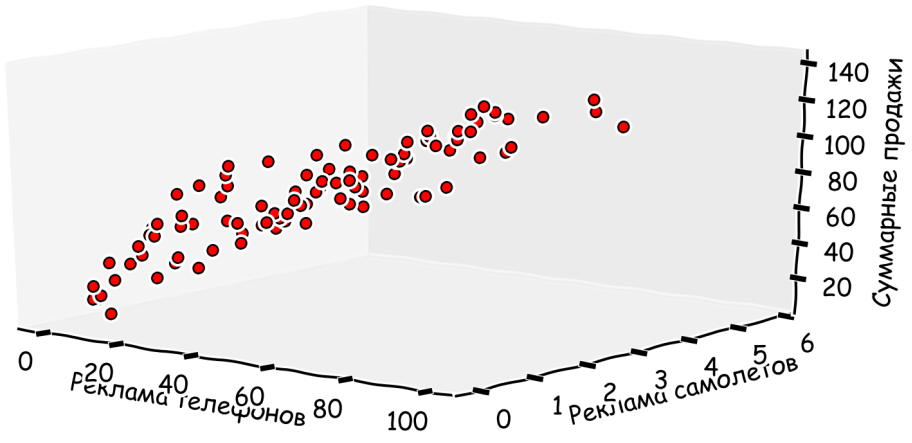

Рис. 11: Зависимость суммарного объема продаж от затрат на рекламу телефонов и на рекламу самолетов

3.3 Статистическая оценка параметров множественной линейной регрессии

В данном пункте аналогично тому, как было сделано в случае простейшей линейной регрессии, можно было бы подробно обсудить всевозможные статические характеристики параметров модели множественной регрессии. Мы не будем этого делать подробно, так как, с одной стороны, вся схема очень похожа, а с другой – доставляет немало технических трудностей. Так что мы ограничимся лишь сводкой важных для практики результатов.

Итак, мы снова будем исходить из следующих важных предположений (первые три из которых часто называют условиями Гаусса-Маркова):

1. Случайные величины (ошибки) $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ независимы и одинаково распределены;

2. Ошибки не носят систематического характера, то есть $\mathsf{E}\varepsilon_i = 0$, $i \in \{1, 2, ..., n\}$;

3. Дисперсии рошибок одинаковы, то есть $\mathsf{D}\varepsilon_i = \sigma^2 > 0$, $i \in \{1, 2, ..., n\}$ (гомоскедастичность);

4. $\varepsilon_i \sim \mathsf{N}_{0,\sigma^2}$.

Не вдаваясь в детали свойств оценок $\widehat{\Theta}$ (предлагаем их сформулировать аналогичным простейшему случаю образом самостоятельно), перейдем сразу к

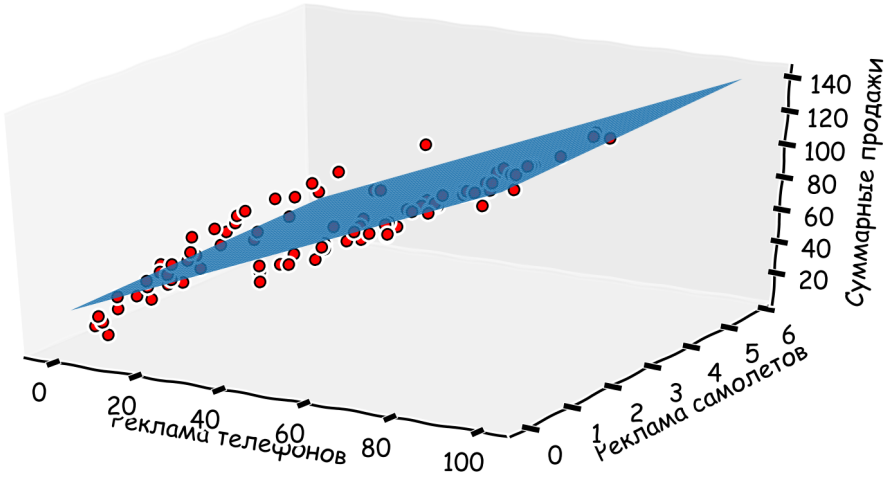

Рис. 12: Зависимость суммарного объема продаж от затрат на рекламу телефонов и на рекламу самолетов и регрессия

практически значимым вопросам – формулам для доверительных интервалов. Для начала получим несмещенную оценку дисперсий σ^2 ошибок ε_i.

Лемма 3.3.1 *В предположении условий 1-4, оценка*

$$\widehat{\sigma_0^2} = \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_{i1} - \widehat{\theta_2}x_{i2} - ... - \widehat{\theta_p}x_{ip})^2 = \frac{1}{n-p-1}|Y - X\widehat{\Theta}|^2$$

является несмещенной оценкой параметра σ^2.

Теперь можно дать явное выражение и для доверительного интервала.

Теорема 3.3.1 (Доверительный интервал для θ_i) *В предположении условий 1 - 4, доверительный интервал уровня доверия* $(1-\varepsilon)$ *для параметра* θ_i *– это интервал*

$$\left(\widehat{\theta_i} - t_{1-\varepsilon/2}\cdot\widehat{\sigma_0}\sqrt{(X^TX)^{-1}_{(i+1)(i+1)}},\ \widehat{\theta_i} + t_{1-\varepsilon/2}\cdot\widehat{\sigma_0}\sqrt{(X^TX)^{-1}_{(i+1)(i+1)}}\right),$$

где $t_{1-\varepsilon/2}$ *– это* $(1-\varepsilon/2)$ *квантиль распределения Стьюдента с* $(n-p-1)$ *степенями свободы, а* $(X^TX)^{-1}_{(i+1)(i+1)}$ *– элемент матрицы* $(X^TX)^{-1}$*, стоящий на пересечении* $(i+1)$*-ой строки и* $(i+1)$*-ого столбца.*

Умея строить доверительные интервалы, ясно, как проверять гипотезы относительно значений коэффициентов регрессии. В частности, как решать задачу о проверке статистической значимости найденных оценок $\widehat{\theta_i}$, $i \in \{0, 1, ..., p\}$.

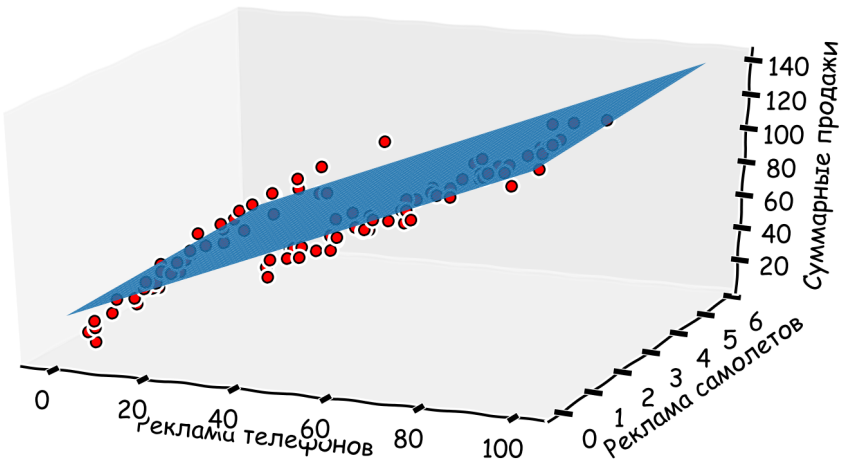

Рис. 13: Зависимость суммарного объема продаж от затрат на рекламу телефонов и на рекламу самолетов и регрессия

Следствие 3.3.2 (О проверке гипотез относительно параметров θ_i) *Пусть проверяется гипотеза $\theta_i = \theta_i^0 \in \mathbb{R}$ против альтернативы $\theta_i \neq \theta_i^0$, то есть*

$$\mathsf{H}_0: \ \theta_i = \theta_i^0$$
$$\mathsf{H}_a: \ \theta_i \neq \theta_i^0.$$

Пусть

$$t = \frac{|\widehat{\theta}_i - \theta_i^0|}{\widehat{\sigma}_0 \sqrt{(X^T X)^{-1}_{(i+1)(i+1)}}},$$

где $(X^T X)^{-1}_{(i+1)(i+1)}$ – элемент матрицы $(X^T X)^{-1}$, стоящий на пересечении $(i+1)$-ой строки и $(i+1)$-ого столбца и $t_{1-\varepsilon/2}$ – квантиль уровня $(1-\varepsilon/2)$ распределения Стьюдента с $(n-p-1)$ степенями свободы, тогда

1. *Если $t > t_{1-\varepsilon/2}$, то гипотеза H_0 отклоняется на уровне значимости ε.*

2. *Если $t \leq t_{1-\varepsilon/2}$, то гипотеза H_0 принимается на уровне значимости ε.*

Примеры применения и интерпретации всех обсужденных методов мы подробно провели в простейшем случае, так что в этом пункте подробно на них не останавливаемся.

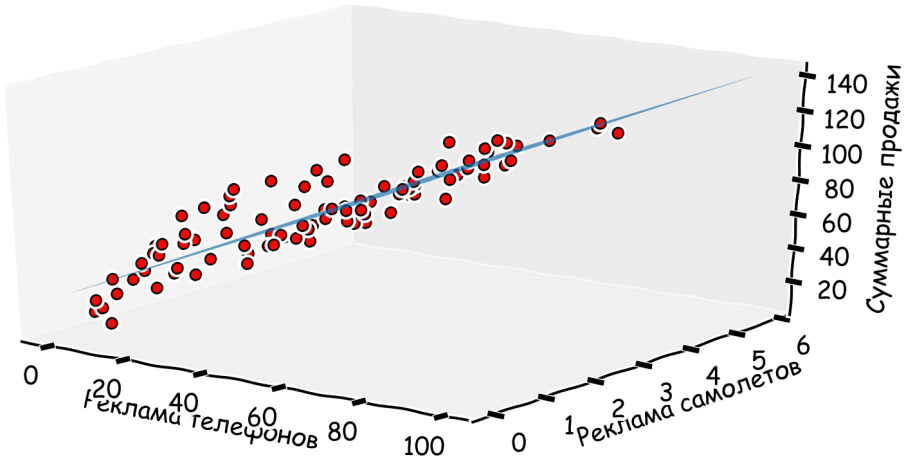

Рис. 14: Зависимость суммарного объема продаж от затрат на рекламу телефонов и на рекламу самолетов и регрессия

3.4 Оценка точности модели множественной регрессии

Как и в случае простейшей регрессии, будем рассматривать RSE и R^2 оценки для нашей модели. Как и ранее, ошибка RSE задается соотношением

$$\mathsf{RSE} = \widehat{\sigma_0} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}\left(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_{i1} - ... - \widehat{\theta_p}x_{ip}\right)^2} =$$

$$= \sqrt{\frac{|Y - X\widehat{\Theta}|^2}{n-p-1}},$$

и показывает «усредненную» сумму квадратов ошибок модели после того, как произведена оценка параметров θ_0, θ_1, ..., θ_p. Аналогично простейшему случаю, вводится в рассмотрение и R^2 статистика:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{\theta_0} - \widehat{\theta_1}x_{i1} - ... - \widehat{\theta_p}x_{ip}\right)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

Снова, аналогично простейшему случаю совершенно понятно, что если построенная идеально соответствует исходным данным, то все слагаемые в числителе последней дроби равны нулю и, тогда и только тогда, $R^2 = 1$, то есть

модель идеальна. Наоборот, если $\widehat{\theta_1} = \widehat{\theta_2} = ... = \widehat{\theta_p} = 0$, то есть модель не зависит от X_1, X_2, ..., X_p, то $\mathsf{R}^2 = 0$, и модель совершенно несостоятельна, так как не отражает никакой зависимости между откликом и предиктором. Как оказывается, значение R^2 лишь увеличивается при добавлении новых предикторов к модели, даже если они очень слабо влияют на отклик.

Можно показать, что взаимосвязь между R^2 и корреляцией r^2 для случая многомерной регрессии такова:

$$\mathsf{R}^2 = r^2\left(Y, \widehat{\theta_0} + \widehat{\theta_1}X_1 + .. + \widehat{\theta_p}X_p\right).$$

В примере множественной регрессии с объемом продаж телефонов и самолетов в зависимости от средств, «вливаемых» в рекламу, коэффициенты таковы: $\mathsf{R}^2 = 0.89$, а $\mathsf{RSE} = 11.04$. Такие значения свидетельствуют о высоком качестве модели.

3.5 Гипотеза о проверке статистической значимости линейной регрессионной модели

Одним из важнейших применений оценки R^2 является проверка статистической значимости построенной модели «в целом», то есть проверка следующего предложения: а есть ли зависимость отклика Y от предикторов X_1, X_2, ..., X_p?

В качестве нулевой гипотезы расморим гипотезу

$$\mathsf{H}_0: \text{ Все параметры } \theta_i \text{ модели равны нулю при } i \in \{1, 2, ..., p\},$$

а в качестве альтернативной гипотезы

$$\mathsf{H}_a: \text{ Хотя бы один из параметров } \theta_i \text{ не равен нулю при } i \in \{1, 2, ..., p\}.$$

Короче это можно записать так:

$$\mathsf{H}_0: \ \theta_1 = \theta_2 = ... = \theta_p = 0,$$

$$\mathsf{H}_1: \ \theta_1^2 + \theta_2^2 + ... + \theta_p^2 \neq 0.$$

Тест проверки гипотез осуществим с помощью F-статистики

$$F = \frac{\mathsf{R}^2}{1 - \mathsf{R}^2}\frac{n - p - 1}{p}.$$

Оказывается, что в условиях 1-4 введенная статистика имеет распределение Фишера с параметрами $(p, n - p - 1)$ (со степенями свободы p и $n - p - 1$). Проверка гипотез осуществляется следующим образом.

Пусть $\varepsilon > 0$ и $f_{1-\varepsilon}$ – квантиль распределения Фишера с праметрами $(p, n - p - 1)$ уровня $(1 - \varepsilon)$.

1. Если $F > f_{1-\varepsilon}$, то гипотеза H_0 отклоняется на уровне значимости α, и, как следствие, построенная модель является статистически значимой.

2. Если $F \leq f_{1-\varepsilon}$, то гипотеза H_0 принимается на уровне значимости α, и, как следствие, построенная модель является статистически незначимой.

4 Немного о полиномиальной регрессии

Как мы уже неоднократно отмечали, линейная регрессия предполагает линейную зависимость между откликом и предикторами. В реальных задачах зависимость может не быть линейной. Оказывается, модель линейной регрессии без существенных усложнений может быть расширена до модели так называемой полиномиальной регрессии. Модель простейшей полиномиальной регрессии имеет вид

$$Y = \theta_0 + \theta_1 X + \theta_2 X^2 + ... + \theta_p X^p + \varepsilon.$$

Важно отметить, что коэффициенты модели могут быть найдены с помощью МНК, описанного выше, ведь перед нами не что иное, как многомерная линейная регрессия, только вместо предикторов взяты степени X:

$$X_1 = X,\ X_2 = X^2, ..., X_p = X^p.$$

Значит, для обсчета такой модели (точнее, для нахождения оценок $\widehat{\theta_0}$, $\widehat{\theta_1}$, ..., $\widehat{\theta_p}$), мы можем использовать аппарат линейной регрессии. Обычно степени X выше четвертой не встречаются, так как полиномиальные кривые могут получаться мало предсказуемой формы.

Значения отклика Y_i по значениям предиктора x_i могут быть получены по правилу

$$Y_i = \widehat{\theta_0} + \widehat{\theta_1} x_i + \widehat{\theta_2} x_i^2 + ... + \widehat{\theta_p} x_i^p, \quad i \in \{1, 2, ..., n\}.$$

Предположим, что нам даны данные, как на рисунке 15. Результат полиномиальной регрессии при различных степенях p можно увидеть на рисунке 16. Синим обозначена классическая линейная регрессия, зеленым – полиномиальная с полиномом второй степени, фиолетовым – полиномиальная с полиномом третьей степени. Видно, что полином третьей степени приближает исходные данные лучше, чем все остальные.

Понятно, что полиномиальная регрессия – лишь частный случай модели, обобщающей модель линейной регрессии. Можно взять какие-то функции φ_1, φ_2, ..., φ_p и построить более общую модель

$$Y = \theta_0 + \theta_1 \varphi_1(X) + \theta_2 \varphi_2(X) + ... + \theta_p \varphi_p(X) + \varepsilon,$$

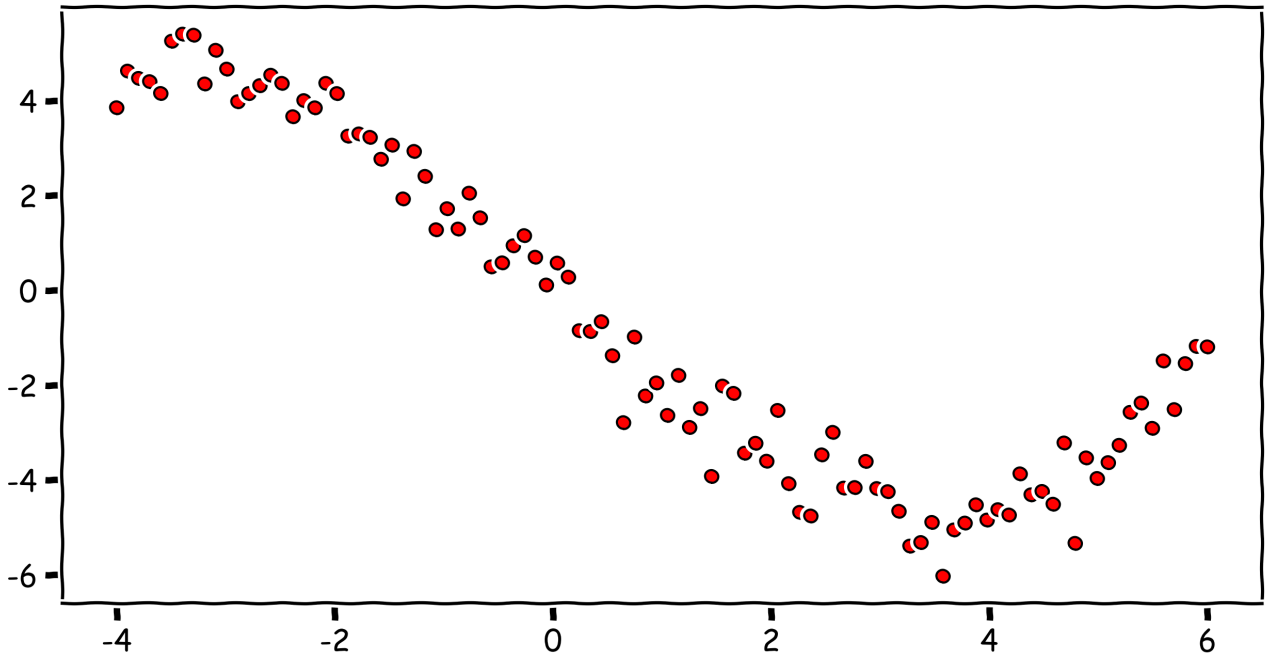

Рис. 15: Набор данных для полиномиальной регрессии

обсчет которой производится, конечно, точно так же, как и в полиномиальном случае. Кроме того, все оценки коэффициентов регрессии проводятся изложенным выше способам.

5 Заключение

В данной лекции мы рассмотрели подходы к решению задачи регрессии. Сам аппарат решения этой задачи достаточно хорошо изучен и подкреплен различными статистическими оценками, однако, как обычно, выбор конкретного набора функций φ_i при построении обобщенной регрессии – задача, которая отдается на откуп исследователю. Удачи!

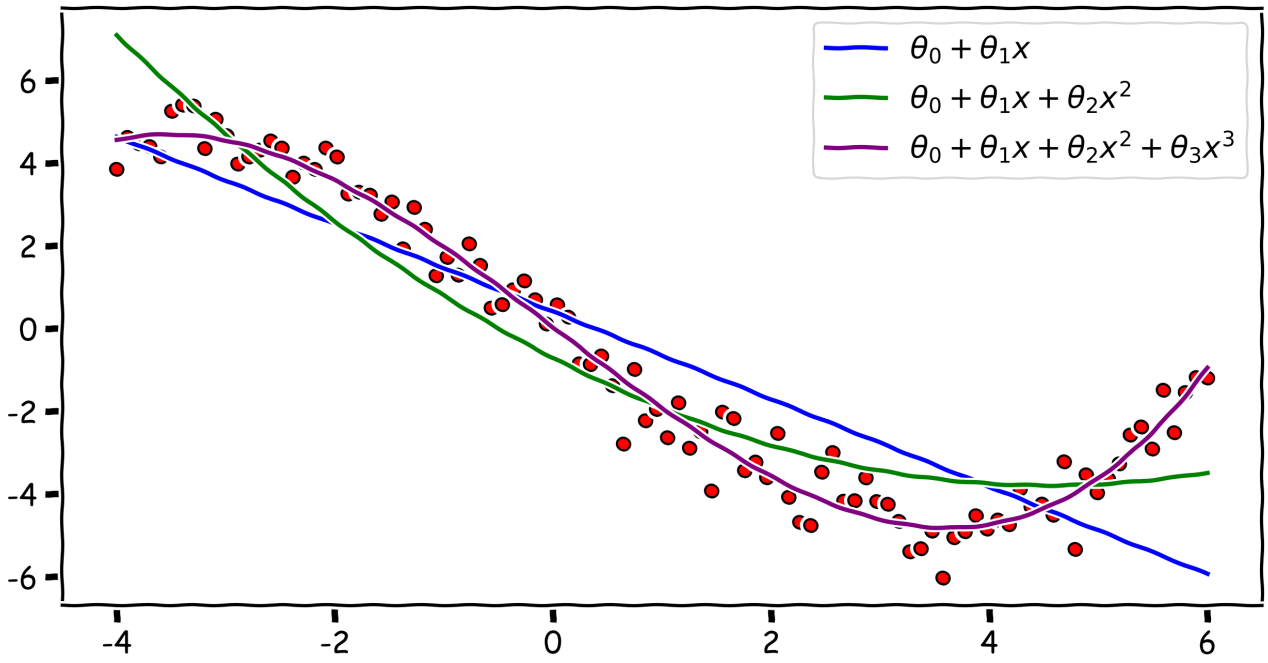

Рис. 16: Полиномиальная регрессии