ЦИФРОВАЯ КУЛЬТУРА

УНИВЕРСИТЕТ ИТМО

Лекция **«Регрессия»**

Бойцев Антон Александрович
Волчек Дмитрий Геннадьевич
Романов Алексей Андреевич

Санкт-Петербург
2019

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.

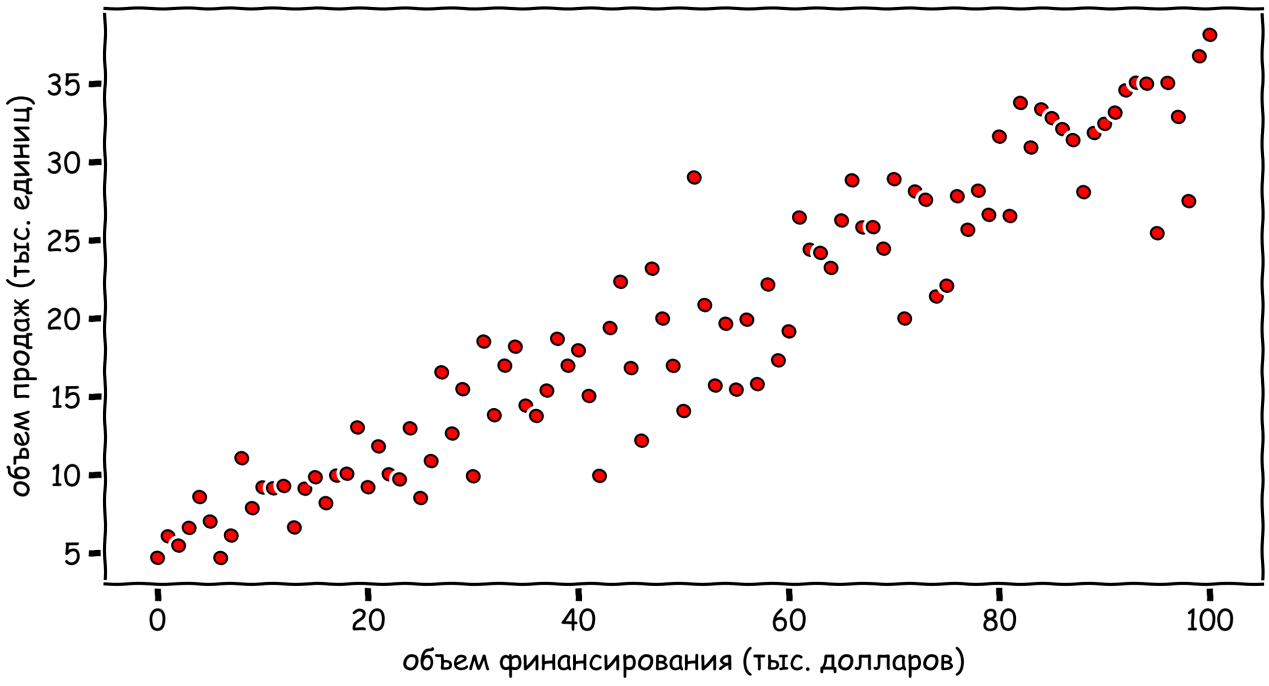

Рис. 1: Зависимость объема продаж мобильных телефонов от затрат на рекламу

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

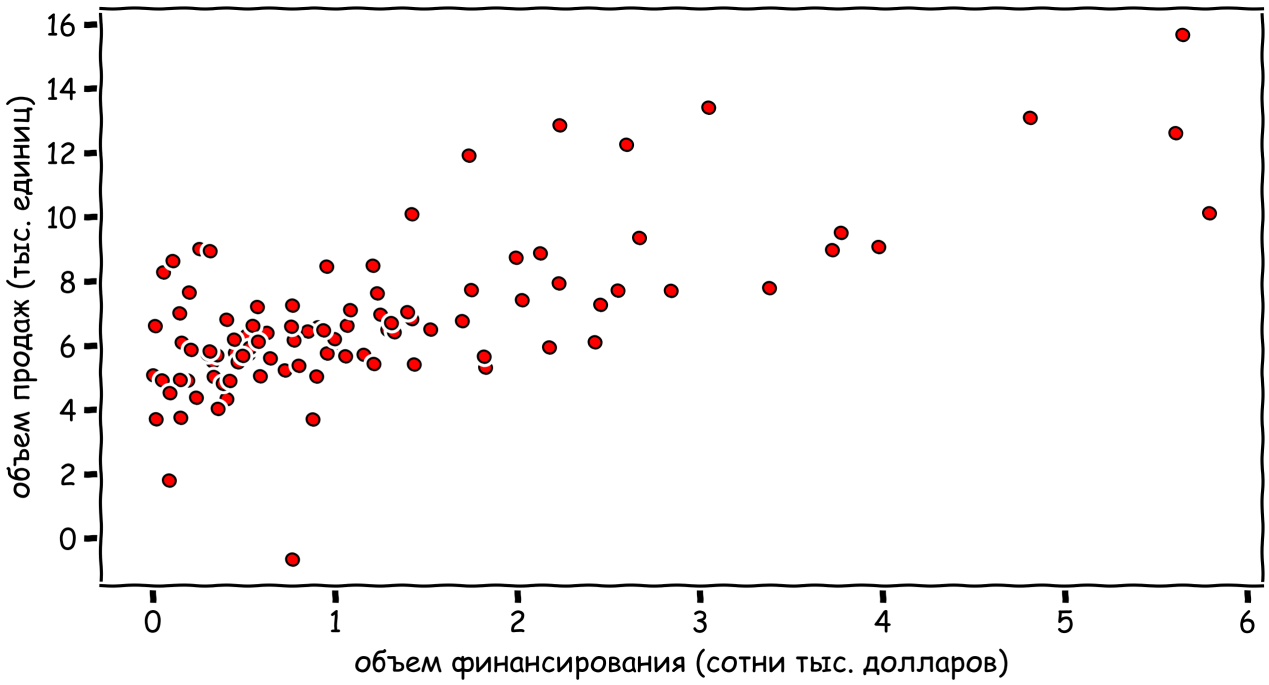

Рис. 2: Зависимость продаж самолетов от объема финансирования

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-

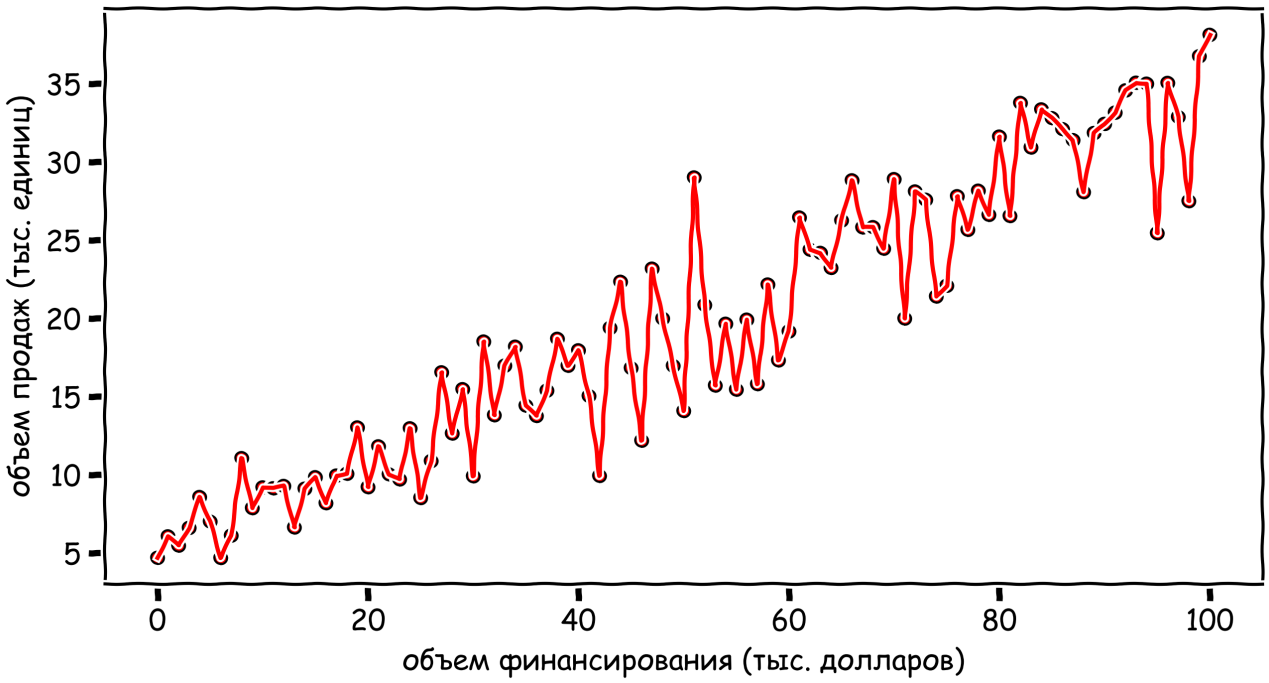

Рис. 3: Наивная зависимость

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 Простейшая модель линейной регрессии

Простейшая линейная регрессия – метод предсказания поведения величины Y (отклика), зная поведение случайной величины X (одного (!) предиктора). Данный метод, что, наверное, ясно и из названия, предполагает линейную зависимость между рассматриваемыми величинами, поэтому модель описывается равенством

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

Есть несомненный плюс данной модели – ее легко трактовать (в отличие от той, что на рисунке 3). Так как в реальности зависимость редко бывает «чисто линейной», скорее лишь приближенно, то на практике правильнее использовать следующую запись:

$$Y \approx \theta_0 + \theta_1 X.$$

Параметры θ_0 и θ_1 в нашей модели неизвестны и для их определения мы будем использовать метод наименьших квадратов (частный случай метода максимального правдоподобия, о котором будет рассказано в следующей лекции). Идея метода широко известна, но мы ее аккуратно и подробно опишем еще раз.

Для начала определим обозначения и поставим математическую задачу. В результате эксперимента у нас есть набор из n пар данных $(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$. В примере с продажами, например, телефонов, первая переменная $X = \{x_1, x_2, ..., x_n\}$ отвечает за объем денежных средств, влитых в рекламу, а вторая переменная $Y = \{y_1, y_2, ..., y_n\}$ – за количество проданных телефонов. Конкретная пара (x_i, y_i) устанавливает соответствие между объемом финансирования x_i, который выделен на рекламу, и количеством продаж телефонов y_i при конкретно этом объеме затрат на рекламу x_i.

Согласно нашему предположению, зависимость приближенно описывается равенством

$$y_i \approx \theta_0 + \theta_1 x_i, \quad i \in \{1, 2, \ldots, n\}.$$

Так как равенство лишь приближенное, а не точное, то на каждой паре $(x_i,\ y_i)$ возникает некоторая ошибка ε_i, которая находится из равенства

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i, \quad i \in \{1, 2, \ldots, n\}.$$

В дальнейшем, найденные нами параметры мы будем обозначать, как θ_0^* и θ_1^*, подразумевая, что они являются оценками истинных значений θ_0 и θ_1. И правда, ведь если исходные данные изменятся, то, скорее всего, найденные нами значения θ_0^* и θ_1^* тоже изменятся, но все также не будут истинными. К

этому моменту неплохо бы понять такую истину: истинных значений, скорее всего, нет вообще!

Но как же найти оценки параметров θ_0 и θ_1? Мы будем пользоваться методом наименьших квадратов (МНК). Этот метод позволяет найти такие оценки θ_0^* и θ_1^* параметров θ_0 и θ_1, что сумма квадратов ошибок $\varepsilon(\theta_0, \theta_1)$ 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.

Ну что, идейная сторона вопроса на этом закончена. С технической же точки зрения перед нами – функция $\varepsilon(\theta_0, \theta_1)$, зависящая от двух переменных θ_0 и θ_1, которую нам требуется минимизировать. 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

$$\theta_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}, \quad \theta_0 = \overline{y} - \theta_1 \overline{x},$$

где $\overline{x}$ и $\overline{y}$ – выборочные средние, то есть

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Конечно, назвать найденные значения θ_0 и θ_1 оценками, а значит и навесить им звезды 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 отбрехаться следующей фразой: функция $\varepsilon(\theta_0, \theta_1)$ выпукла вниз, а значит найденная точка, подозрительная на экстремум, и правда является точкой минимума.

Теперь можно смело записать найденные оценки:

$$\theta_1^* = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}, \quad \theta_0^* = \overline{y} - \theta_1 \overline{x},$$

где $\overline{x}$ и $\overline{y}$ – выборочные средние, то есть

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Итак, на основе данных по мобильным телефонам, получаем значения $\theta_0^* \approx 4.88$, $\theta_1^* \approx 0.30$ (более подробно все вычисления мы объясним на следующем примере, где не так много исходных данных, и вычисления более компактны). В итоге, функция

$$y = 4.88 + 0.30x$$

и есть искомая функция, дающая модель простейшей линейной регрессии. Построим ее график, он изображен на рисунке 4 синим цветом.

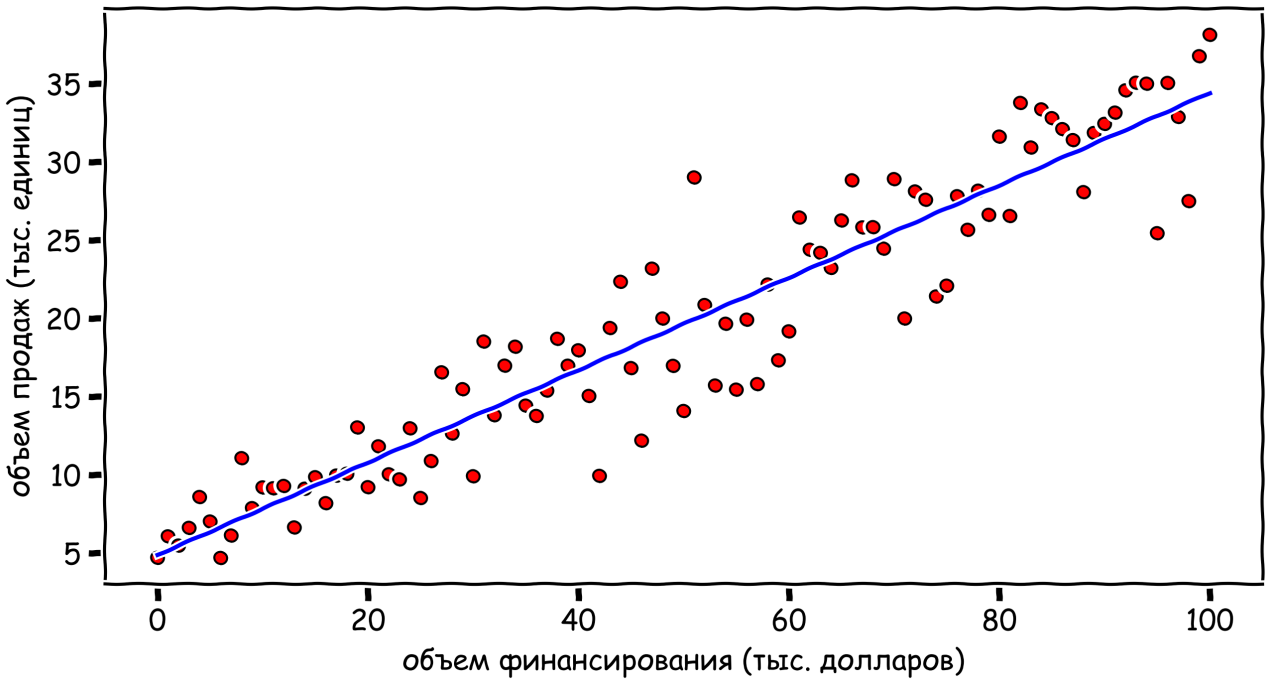

Рис. 4: Зависимость объема продаж мобильных телефонов от затрат на рекламу и регрессия

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 ошибки ε_i (сумму квадратов которых, мы минимизировали), которые показывают отклонение нашей модели от реальных данных, рисунок 5.

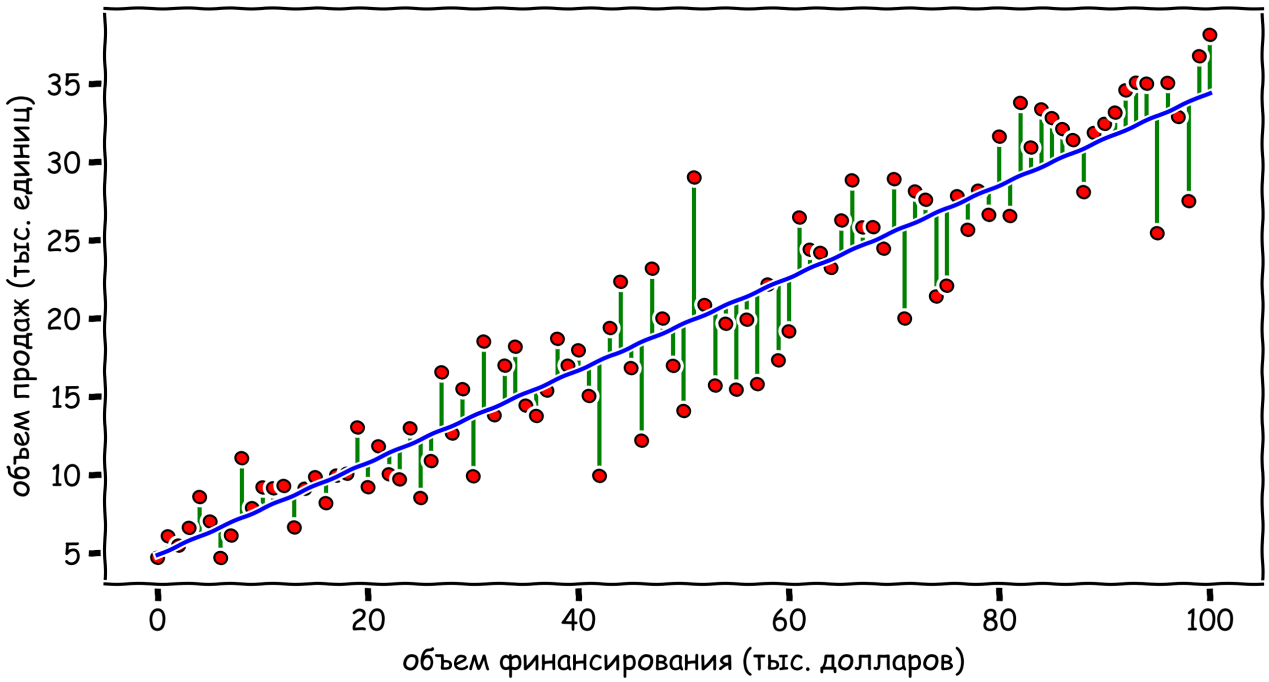

Рис. 5: Зависимость объема продаж мобильных телефонов от затрат на рекламу, регрессия и ошибки

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

$$y = 5.13 + 1.34x$$

и есть искомая функция, дающая модель простейшей линейной регрессии. Построим ее график, он изображен на рисунке 6 синим цветом. Детальное обсуждение точности данной модели, как и модели, полученной ранее, проведем чуть позже.

После увиденного, еще раз подчеркнем, что, найдя оценки θ_0^* и θ_1^* предсказание ищется, согласно формуле

$$Y = \theta_0^* + \theta_1^* X.$$

Без дополнительных теоретических пояснений отметим, что для того, чтобы МНК был подходящим методом требуется, чтобы:

1. Математическое ожидание ошибок было равно нулю;

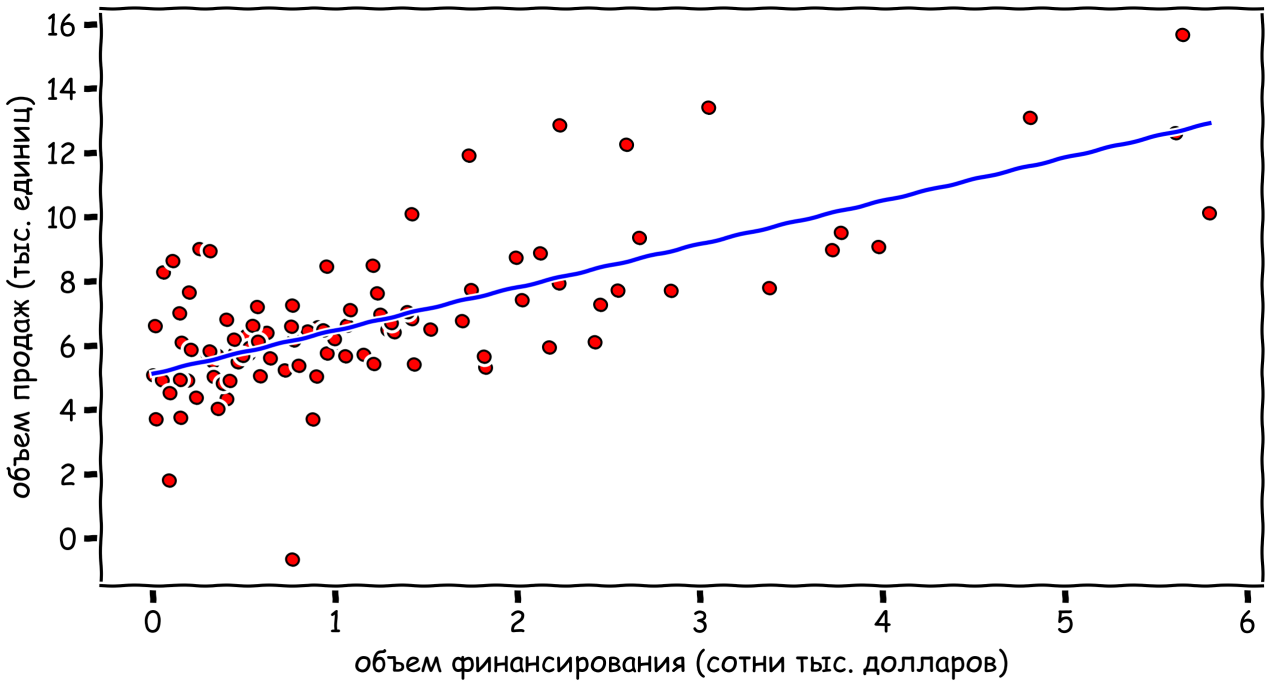

Рис. 6: Зависимость объема продаж самолетов от затрат на рекламу и регрессия

2. Дисперсия ошибок была постоянной величиной;

3. Отсутствует корреляция ошибок, ковариация равна нулю.

Обычно предполагают, что ошибка имеет нормальное распределение с математическим ожиданием, равным нулю, а дисперсия ошибки постоянна и не зависит от входных переменных.

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.

№ 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 предиктором, а что откликом. В качестве предиктора X 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 7. По горизонтальной оси отложены иксы, а по вертикальной – игреки.

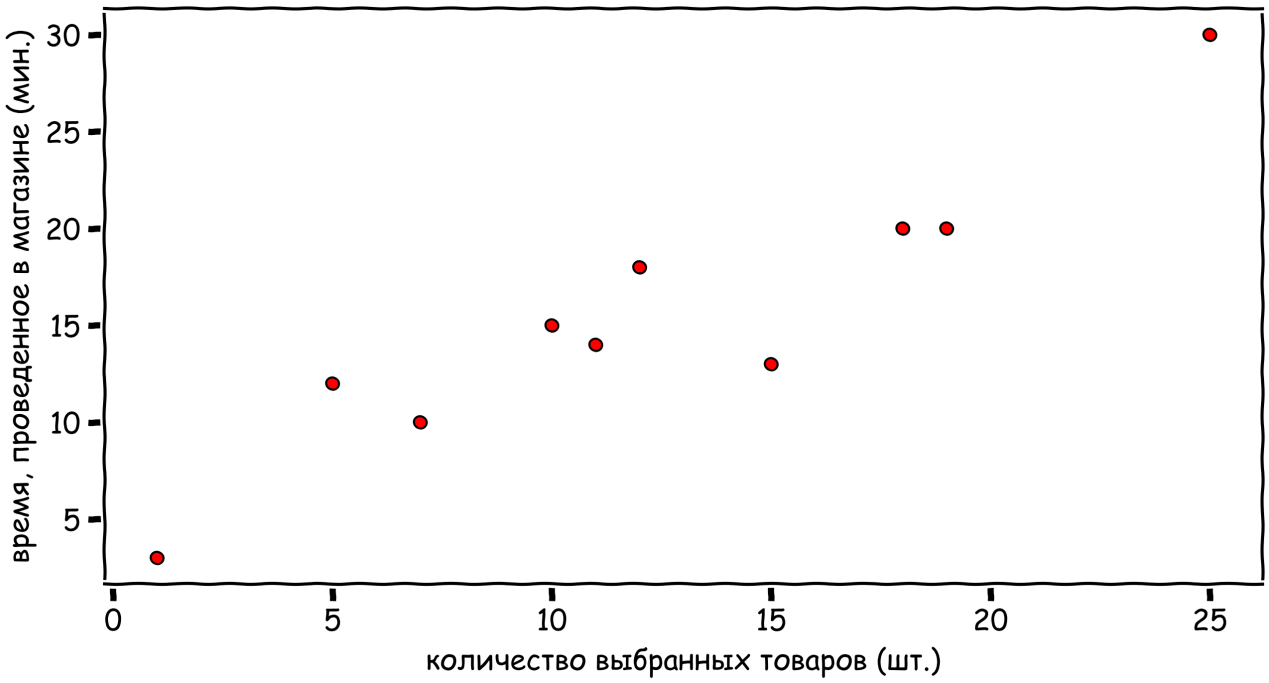

Рис. 7: Зависимость времени, проведенного в магазине, от количества выбранных товаров

Так как в нашем случае $n = 10$, то формулы для θ_0^* и θ_1^* примут следующий вид:

$$\theta_1^* = \frac{\sum\limits_{i=1}^{10}(x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=1}^{10}(x_i - \overline{x})^2}, \quad \theta_0^* = \overline{y} - \theta_1 \overline{x},$$

а для выборочных средних $\overline{x}$ и $\overline{y}$:

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i, \quad \overline{y} = \frac{1}{10}\sum_{i=1}^{10} y_i.$$

Начнем с вычисления последних, итак

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

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

Теперь мы можем вычислить θ_1^*:

$$\theta_1^* = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + ... + (x_{10} - \overline{x})(y_{10} - \overline{y})}{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... + (x_{10} - \overline{x})^2} =$$

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

Ну а тогда

$$\theta_0^* \approx 15.5 - 0.93 \cdot 12.3 \approx 4.06.$$

В реальных подсчетах значения лучше не округлять, и подставлять для расчета θ_0^* значение θ_1^* с как можно большим числом знаков после запятой. Мы округлили θ_1^* и нашли приближенное значение θ_0^* для наглядности.

Итак, уравнение линейной регрессии имеет следующий вид:

$$y = 4.06 + 0.93x.$$

Построим получившуюся прямую, результат можно видеть на рисунке 8. Вер-

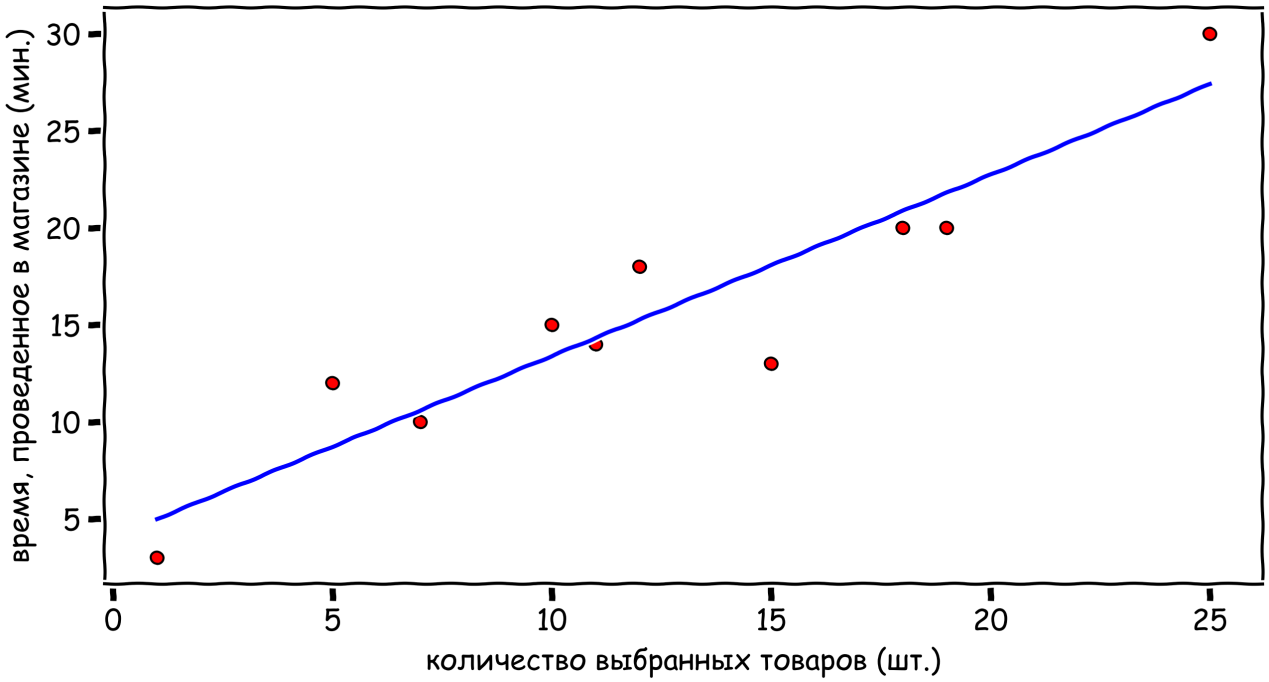

Рис. 8: Зависимость времени, проведенного в магазине, от количества выбранных товаров и регрессия

немся 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 $y = 4.06 + 0.93x$ при $x = 27$, то есть

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

а значит потребуется чуть больше, чем 29 минут. Иллюстрация прогноза приведена на рисунке 9.

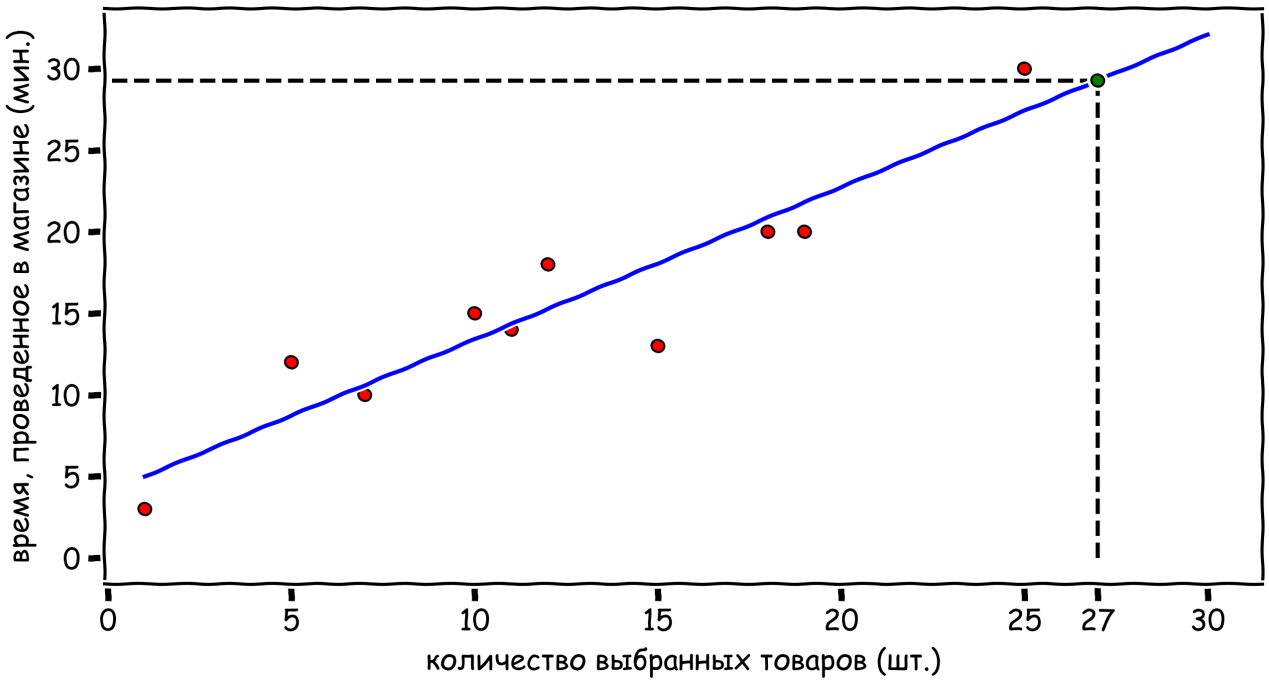

Рис. 9: Зависимость времени, проведенного в магазине, от количества выбранных товаров, регрессия и предсказание

1.4 Построение доверительных интервалов

Все, что описано выше, было сделано, основываясь на одной выборке. В то же время, проведя исследование еще раз, мы можем получить отличающиеся данные, а значит новые оценки параметров θ_0 и θ_1 не будут в точности равны нашим (например, как мы уже говорили, продажи мобильных телефонов возрастают ближе к празднику нового года, и падают ближе к сезону отпусков). Однако если нам позволено проводить эксперимент много раз, то взяв в качестве параметров модели среднее полученных оценок, новая модель будет еще лучше описывать реальное положение дел.

Оценим «разброс» среди возможных оценок θ_0^* и θ_1^*. Стандартные ошибки (standard error) оценок θ_0^* и θ_1^* могут быть найдены по формулам

$$\mathsf{SE}(\theta_0^*) = \sqrt{\frac{\sum\limits_{i=1}^{n}(y_i - \overline{y})^2}{n-2}} \cdot \sqrt{\frac{1}{n} + \frac{\overline{x}}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}},$$

$$\mathsf{SE}(\theta_1^*) = \sqrt{\frac{\sum\limits_{i=1}^{n}(y_i - \overline{y})^2}{n-2}} \cdot \sqrt{\frac{1}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}}$$

В случае данных с мобильными телефонами величины $\mathsf{SE}(\theta_0^*)$ и $\mathsf{SE}(\theta_1^*)$ соответственно равны 0.94 и 0.03, а в случае данных с самолетами – 0.32 и 0.19.

На основе стандартных ошибок можно построить так называемый доверительный интервал. Пусть $\varepsilon > 0$. Доверительный интервал $(\theta^-,\ \theta^+)$ уровня доверия $1-\varepsilon$ – это интервал, в который с вероятностью $1-\varepsilon$ попадет реальное значение параметра. Обычно рассматривают значения $\varepsilon = 0.1$, $\varepsilon = 0.05$ или $\varepsilon = 0.01$. В линейной регрессии $(1-\varepsilon)$ доверительный интервал для параметра θ_0 может быть записан, как

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

где t – это $(1-\varepsilon/2)$ квантиль распределения Стьюдента с $(n-2)$ степенями свободы. Эти значения ищутся в таблице, таблица приложена в дополнительных материалах к лекции.

Аналогично, доверительный интервал уровня доверия $(1-\varepsilon)$ для θ_1 может быть записан в виде

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

где t – опять же $1-\varepsilon/2$ квантиль распределения Стьюдента с $(n-2)$ степенями свободы.

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

$$\left(\theta_0^-,\ \theta_0^+\right) = (3.31,\ 6.44),$$

а для θ_1 имеет вид

$$\left(\theta_1^-,\ \theta_1^+\right) = (0.24,\ 0.35).$$

Подробный пример расчета мы увидим чуть позже, а сейчас давайте вдумаемся в смысл полученных интервалов. Можно заметить, что интервал для значения θ_0 получился намного шире, чем для значения θ_1, что, скорее всего объясняется довольно большим разбросом в данных.

На основе полученных интервалов можно сделать вывод, что при отсутствии рекламы продажи, в среднем, упадут до $3.31-6.44$ тысяч единиц. При этом, за каждую потраченную на рекламу тысячу долларов объем продаж в среднем увеличится на $0.24-0.35$ тысяч единиц.

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

$$\left(\theta_0^-,\ \theta_0^+\right) = (4.59,\ 5.66)$$

а для θ_1 имеет вид

$$\left(\theta_1^-,\ \theta_1^+\right) = (1.02,\ 1.66).$$

Проанализировав эти результаты, можно сделать вывод, что при отсутствии рекламы продажи, в среднем, упадут до $4.59 - 5.66$ тысяч единиц. При этом за каждую потраченную на рекламу сотню тысяч долларов объем продаж в среднем увеличится на $1.02 - 1.66$ тысячу единиц.

1.5 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:

$$\mathsf{SE}(\theta_0^*) = \sqrt{\frac{\sum\limits_{i=1}^{10}(y_i - \overline{y})^2}{10-2}} \cdot \sqrt{\frac{1}{10} + \frac{\overline{x}}{\sum\limits_{i=1}^{10}(x_i - \overline{x})^2}},$$

$$\mathsf{SE}(\theta_1^*) = \sqrt{\frac{\sum\limits_{i=1}^{10}(y_i - \overline{y})^2}{10-2}} \cdot \sqrt{\frac{1}{\sum\limits_{i=1}^{10}(x_i - \overline{x})^2}},$$

выборочные средние $\overline{x}$ и $\overline{y}$ вычислены нами ранее и равны 12.3 и 15.5, соответственно.

В самом начале разбора примера мы достаточно подробно описали, как работать со знаком суммы и что есть что, так что здесь ограничимся только финальными результатами. В то же время 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 $\mathsf{SE}(\theta_0^*) = 2.72$ и $\mathsf{SE}(\theta_1^*) = 0.36$, то

$$\left(\theta_0^-,\ \theta_0^+\right) = (-0.98,\ 9.1)$$

и

$$\left(\theta_1^-,\ \theta_1^+\right) = (0.27,\ 1.60)\,.$$

Что же значат эти интервалы? Рассмотрим подробнее второй. 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 0.27 минут до 1.6 минут. Первый же интервал показывает, что, проведя в магазине ноль минут, можно в среднем купить 5 товаров. 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 данный вопрос снимается.

1.6 Проверка гипотез

Стандартные ошибки также используются в так называемой задаче проверки гипотез. Одна из самых часто проверяемых гипотез такова

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

Альтернативная ей гипотеза такова

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

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

$$\mathsf{H}_0: \ \theta_1 = 0,$$
$$\mathsf{H}_1: \ \theta_1 \neq 0.$$

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

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

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.

2. Если $t_{1-\varepsilon/2} \geq t$, то гипотеза H_0 принимается.

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 $t = 9.39$, 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 $t = 6.98$, что снова больше значения $t_{1-\varepsilon/2}$ из таблицы, значит гипотеза H_0 отклоняется и принимается альтернативная гипотеза H_1.

1.7 Проверка гипотез для примера

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

$$t = \frac{\theta_1^*}{\mathsf{SE}(\theta_1^*)} = \frac{0.93}{0.36} \approx 2.58.$$

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

1.8 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: среднее квадратическое 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. RSE – оценка среднего квадратичного отклонения ошибки ε. Грубо говоря, показывается насколько ответ модели отличается от «настоящей» линии регрессии. RSE может быть вычислена по формуле

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

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

$$\mathsf{R}^2 = 1 - \frac{\sum\limits_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)^2}{\sum\limits_{i=1}^{n}(y_i - \overline{y})^2}.$$

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 и Y определяется, как

$$r(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

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

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.

1.9 Оценка точности модели для примера

Для нашего примера формулы переписываются в следующем виде:

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

$$\mathsf{R}^2 = 1 - \frac{\sum_{i=1}^{10}(y_i - 4.06 - 0.93 \cdot x_i)^2}{\sum_{i=1}^{10}(y_i - \overline{y})^2},$$

выборочно среднее $\overline{y}$ вычислено нами ранее и равно 15.5.

В нашем примере со временем, проведенным в магазине, параметры таковы:

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

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

2 Многомерная линейная регрессия

Простейшая линейная регрессия показывает, как предсказать значение 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?

2.1 Формулировка задачи

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

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

где $X_1, X_2, ..., X_p$ – входные данные (предикторы) по которым мы пытаемся определить переменную Y (отклик). Конечно, вместо знака $=$ стоит снова писать $\approx$. Ясно, что относительно данной нам выборки, состоящей из элементов

$$(x_{11}, x_{12}, ..., x_{1p}, y_1),\ (x_{21}, x_{22}, ..., x_{2p}, y_2),\ ...,\ (x_{n1}, x_{n2}, ..., x_{np}, y_n)$$

мы получаем значение y_i по значениям $x_{i1}, x_{i2}, ..., x_{ip}$, а значит

$$y_i \approx \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + ... + \theta_p x_{ip}.$$

Если говорить точнее, то

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

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

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

Для нахождения оценок неизвестных параметров аналогично тому, что было сделано в одномерном случае, применим МНК. Тем самым оценки $\theta_0^*, \theta_1^*, ..., \theta_p^*$

коэффициентов $\theta_0, \theta_1, ..., \theta_p$ находятся из решения задачи минимизации функции $\varepsilon(\theta_0, \theta_1, ..., \theta_p)$, зависящей уже от p переменных

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

Не будем вдаваться в детали получения оценок, так как схема действий не меняется, лишь становится более громоздкой. При $\det\left(X^T X\right) > 0$ получим замкнутое выражение

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

где индекс T над матрицей означает, ее транспонирование. Полезно выписать формулы для оценок явно в случае, когда $p = 2, 3$. Зная оценки коэффициентов модели предсказание может быть сделано в соответствии с формулой

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

Когда мы ведем разговор о многомерной регрессии, то возникают вопросы следующего характера:

1. Является ли хотя бы один из предсказателей $X_1, X_2, ..., X_p$ полезным для отклика?

2. Все ли предикторы помогают найти Y, или только какая-то их часть?

3. Насколько хорошо модель описывает данные?

На все эти вопросы, аналогично тому, как было сделано в одномерном случае, мы ответим дальше. Второй вопрос заслуживает отдельного внимания и будет обсуждаться позднее.

Обратимся к примеру построения многомерной регрессии. 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 10. Ясно, что в нашем случае количество предикторов равно двум, а значит модель имеет следующий вид:

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

Найдем оценки θ_0^*, θ_1^* и θ_2^* неизвестных коэффициентов θ_0, θ_1 и θ_2. Используя вышеприведенные формулы, имеем

$$\theta_0^* = 27.20, \quad \theta_1^* = 1.08, \quad \theta_2^* = 0.86,$$

и предсказание осуществляется по формуле

$$y = 27.20 + 1.08 \cdot x_1 + 0.86 \cdot x_2.$$

Попытки изобразить плоскость и разброс данных с разных ракурсов приведены на рисунках 11, 12 и 13:

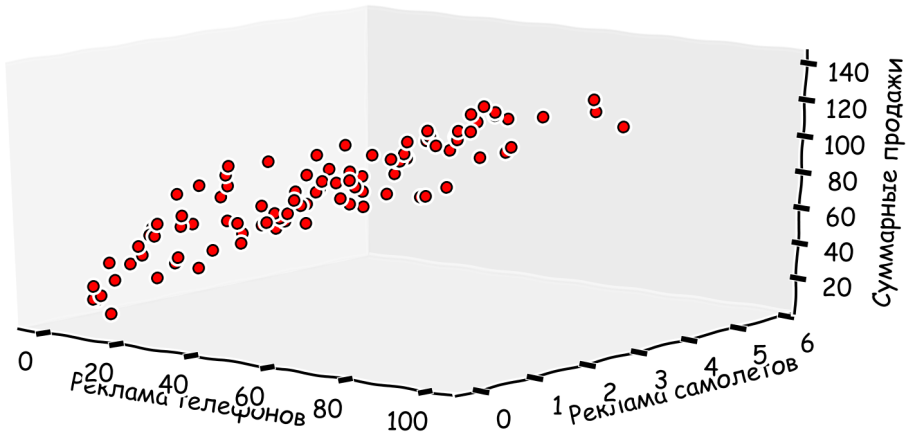

Рис. 10: Зависимость суммарного объема продаж от затрат на рекламу телефонов и на рекламу самолетов

2.2 Оценка предсказателей

Аналогично тому, как мы поступали в одномерном случае, резонно поставить задачу о проверке гипотез. В качестве нулевой гипотезы возьмем

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

а в качестве альтернативной гипотезы

$$\mathsf{H}_1: \text{ Хотя бы один из параметров } \theta_i \text{ не равен нулю при } i \in \{1, 2, ..., p\}.$$

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

$$F = \frac{n-p-1}{p} \cdot \frac{\sum\limits_{i=1}^{n} (y_i - \overline{y})^2 - \sum\limits_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_{i1} - ... - \theta_p x_{ip})^2}{\sum\limits_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_{i1} - ... - \theta_p x_{ip})^2}.$$

Если условия применимости линейной модели не нарушены, то отсутствие взаимосвязи между откликом и предикторами выражено в том, что значение

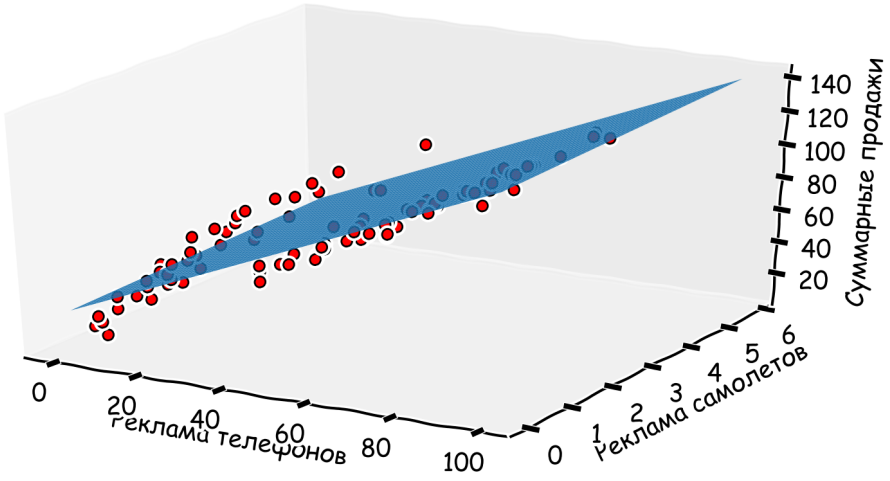

Рис. 11: Зависимость суммарного объема продаж от затрат на рекламу телефонов и на рекламу самолетов и регрессия

F близко к единице. В этом случае справедлива гипотеза H_0. Если же значение больше единицы, то принимается гипотеза H_1. Более точно проверка гипотез проводится при помощи таблиц распределения Фишера аналогично тому, как было проведено в одномерном случае.

В нашем случае значение F-статистики равно 411.82, что свидетельствует о том, что отклик на предикторы установлен.

2.3 Оценка модели

Как и в случае простейшей регрессии, будем рассматривать RSE и R^2 оценки для нашей модели. Можно показать, что взаимосвязь между R^2 и r^2 для случая многомерной регрессии такова: $R^2 = r^2\left(Y, \theta_0 + \theta_1 X_1 + .. + \theta_p X_p\right)$ Значение коэффициента R^2, близкое к 1 показывает, что отклик зависит от предикторов. Как оказывается, значение R^2 лишь увеличивается при добавлении новых предикторов к модели, даже если они очень слабо влияют на отклик.

Для вычисления RSE формула также несколько скорректируется:

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

В случае $p = 1$, то есть в случае простейшей линейной регрессии, написанная

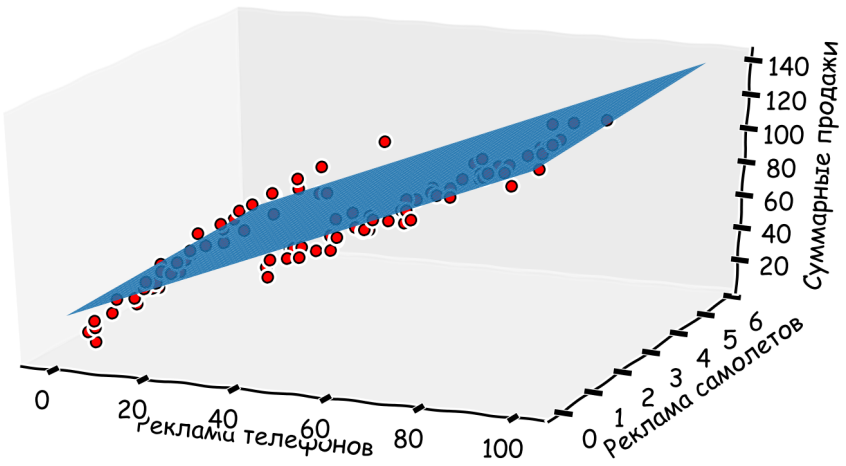

Рис. 12: Зависимость суммарного объема продаж от затрат на рекламу телефонов и на рекламу самолетов и регрессия

формула для RSE как раз-таки превращается в ту, что мы давали ранее:

$$\mathsf{RSE} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_{i1} - ... - \theta_p x_{ip})^2}.$$

В добавление к параметрам RSE и R^2, бывает удобно построить графики зависимостей и сделать какие-то выводы исходя из увиденного. В нашем примере $\mathsf{R}^2 = 0.89$, а $\mathsf{RSE} = 11.04$, что свидетельствует о высоком качестве модели.

2.4 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

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

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:

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

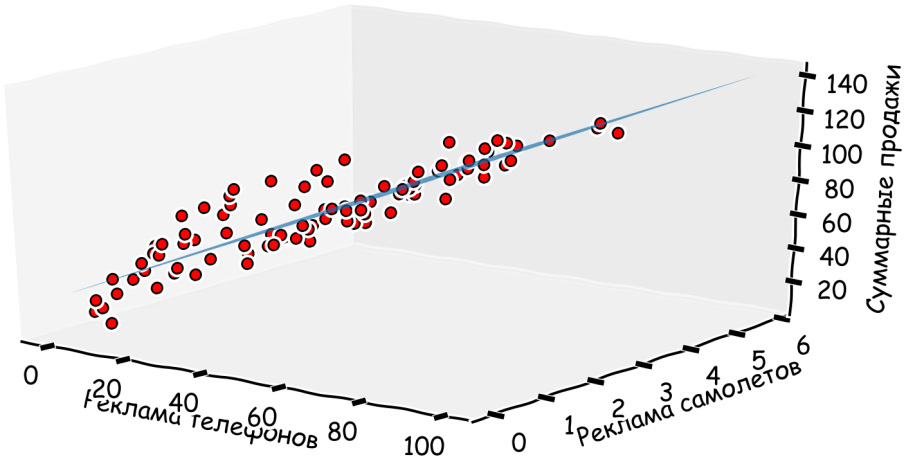

Рис. 13: Зависимость суммарного объема продаж от затрат на рекламу телефонов и на рекламу самолетов и регрессия

Значит, для обсчета такой модели (точнее, для нахождения оценок θ_0^*, θ_1^*, ..., θ_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

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

Предположим, что нам даны данные, как на рисунке 14

Результат полиномиальной регрессии при различных степенях p можно увидеть на рисунке 15.

Синим обозначена классическая линейная регрессия, зеленым – полиномиальная с полиномом второй степени, фиолетовым – полиномиальная полиномом третьей степени. Видно, что полином третьей степени приближает исходные данные лучше, чем все остальные.

2.5 Пара слов про Ридж и LASSO регрессию

В этом пункте мы возвращаемся к вопросу, заданному в многомерной регрессии: все ли предикторы активно задействованы в поиске отклика Y или только их часть? Если незадействованные переменные исключить, то модель будет проще интерпретировать. Методы Ридж и LASSO регрессии представляют собой некоторую альтернативу технике «выбора подмножества», на которой мы подробно не останавливаемся, для уменьшения числа предикторов.

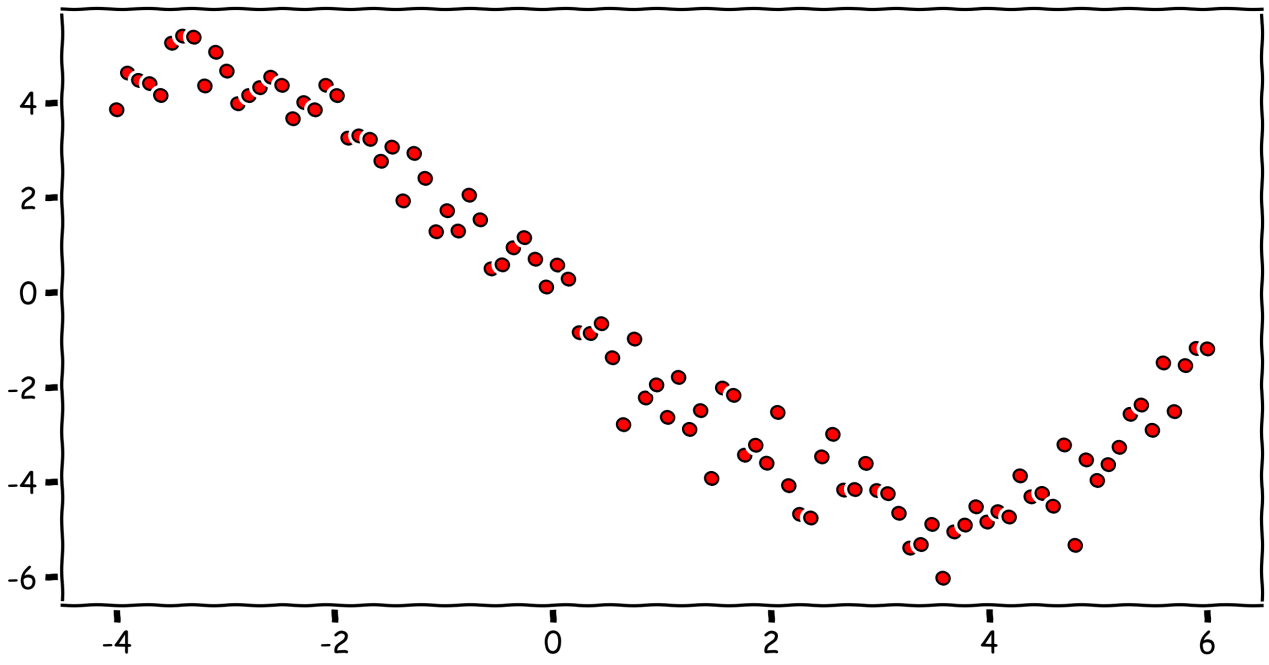

Рис. 14: Набор данных для полиномиальной регрессии

В результате применения этих методов коэффициент (вес) при некоторых предикторах нашей модели приближается к нулю (или обнуляется вовсе).

Из многомерной регрессии мы помним, что ставится задача минимизации выражения

$$\sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_{i1} - \theta_2 x_{i2} - ... - \theta_p x_{ip})^2 = \sum_{i=1}^{n} \left(y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij} \right)^2 .$$

В ридж-регрессии (или еще ее называют гребневой регрессией) минимизируется выражение

$$\sum_{i=1}^{n} \left(y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \theta_j^2,$$

где параметр $\lambda \geq 0$ определяется отдельно. Метод наименьших квадратов находит такие оценки θ_i^* коэффициенты θ_i, что первая сумма минимальна. Однако вторая сумма, так как она содержит квадраты величин, минимальна тогда, когда минимальны по абсолютной величине θ_i^*. Коэффициент λ контролирует роль каждой из двух сумм в оценке коэффициентов θ_i. Например, при $\lambda = 0$ вторая сумма не вносит никакой вклад, однако при $\lambda \to +\infty$ ее вклад очень велик, и коэффициенты в ридж-регрессии будут стремительно идти к нулю.

Сама модель напоминает метод множителей Лагранжа поиска условного экстремума функции многих переменных. Для поиска параметра λ часто

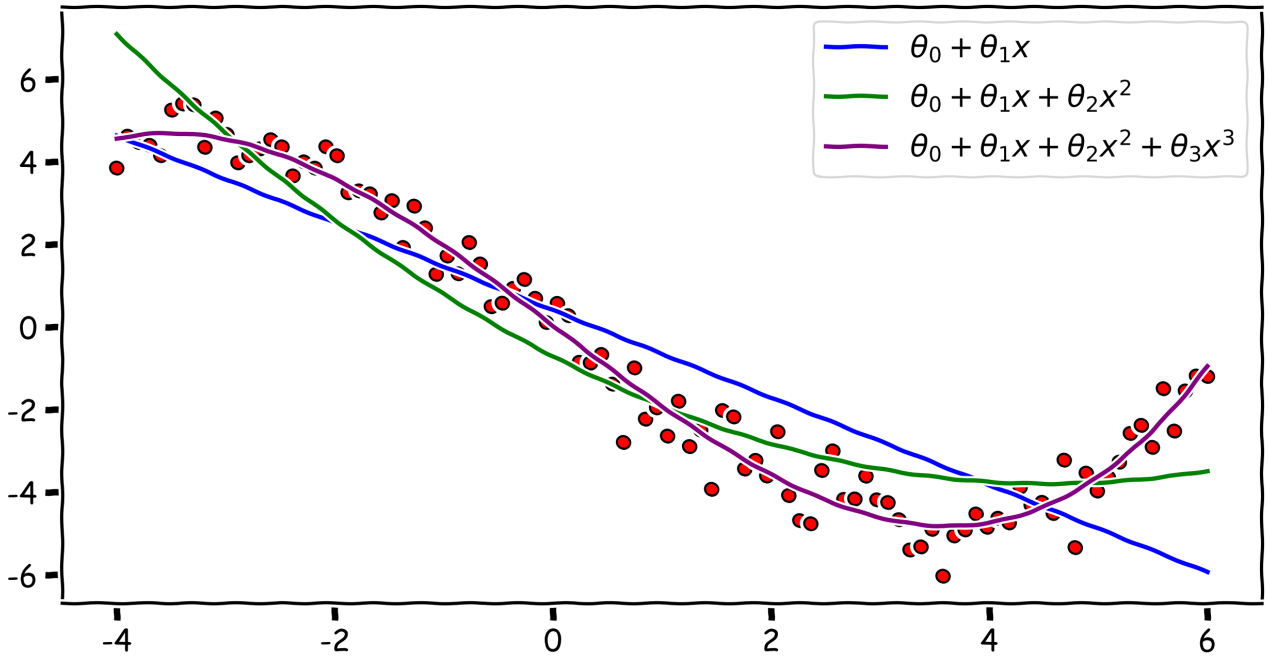

Рис. 15: Полиномиальная регрессии

используется метод кросс-валидации, о котором будет сказано в следующих лекциях.

Метод ридж-регрессии имеет один принципиальный недостаток. Дело в том, что модель все равно включает в себя все предикторы, даже если коэффициенты перед ними становятся чрезвычайно малы. Это может не быть проблемой в точности моделирования и предсказания отклика, но может стать серьезной проблемой при интерпретации модели, особенно если p велико. Метод LASSO регрессии достаточно новый метод, который позволяет решить возникающую трудность. Коэффициенты в методе минимизируют выражение

$$\sum_{i=1}^{n} \left(y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\theta_j|.$$

Отличие заключается в том, что во второй суммы каждое слагаемое θ_j^2 заменено на $|\theta_j|$. Метод регрессии LASSO тоже старается сделать коэффициенты ближе к нулю, но позволяет им быть в точности нулем при достаточно больших значениях параметра λ. Как результат, модели, построенные с помощью метода LASSO, часто бывает проще интерпретировать, чем модели, построенные с помощью ридж-регрессии. Для нахождения удобного параметра λ опять же часто используется метод кросс-валидации.