Тема 2. ПАРНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

1. Модель парной линейной регрессии
2. Регрессия методом наименьших квадратов
3. Интерпретация уравнения регрессии
4. Качество оценивания: коэффициент R^2

1. Модель парной линейной регрессии

Коэффициент корреляции может показывать, что две переменные связаны друг с другом, однако он не дает представления о том, каким образом они связаны. Рассмотрим более подробно те случаи, для которых мы хотим предположить, что одна переменная, называемая обычно **зависимой переменной,** определяется другими переменными, называемыми **объясняющими переменными** (а также **независимыми переменными** или **регрессорами**). Предполагаемая математическая зависимость, связывающая эти переменные, называется **моделью регрессии.** Если в модели присутствует только один регрессор, модель называется **моделью парной регрессии.** Если в модели присутствует два или более регрессора, то она называется **моделью множественной регрессии.**

Не следует ожидать получения точного соотношения между какими-либо двумя экономическими показателями, за исключением тех случаев, когда оно существует по определению. В учебниках по экономической теории эта проблема обычно решается путем приведения соотношения, как если бы оно было точным, и предупреждения читателя о том, что это аппроксимация. В статистическом анализе факт неточности соотношения признается путем явного включения в него случайного фактора, описываемого **случайным (остаточным) членом.**

Начнем с рассмотрения простейшей модели:

$$Y_i = \beta_1 + \beta_2 x_i + u_i \qquad (1.1)$$

Величина Y_i, значение зависимой переменной в наблюдении i, состоит из двух составляющих: 1) неслучайной составляющей $\beta_1 + \beta_2 x_i$, где β_1 и β_2 - это постоянные величины, называемые **параметрами** уравнения; а X − значение объясняющей переменной в наблюдении i, и 2) случайного члена u_i.

На рис. 1.1 показано, как комбинация этих двух составляющих определяет величину Y. Показатели $X_1 X_2 X_3$ и X_4 — это четыре гипоте-

тических значения объясняющей переменной. Если бы соотношение между Y и X было точным, то соответствующие значения *Y* были бы представлены точками Q_1 – Q_4 на прямой. Наличие случайного члена приводит к тому, что в действительности значение Y получается другим. Предполагалось, что случайный член положителен в первом и четвертом наблюдениях и отрицателен в двух других, поэтому если отметить на графике реальные значения *Y* при соответствующих значениях *X*, то мы получим точки P_1 - P_4.

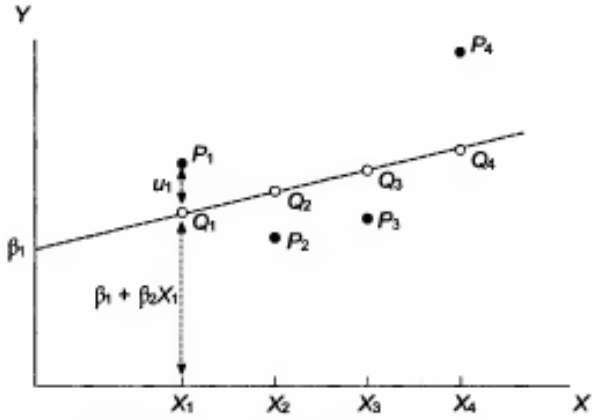

Рисунок 1.1. Истинная зависимость между *Y* и *X*

Следует подчеркнуть, что точки *P* — это все, что вы можете видеть на рис.1.1 на практике. Фактические значения β_1и β_2 и, следовательно, положения точек *Q* неизвестны, так же как и фактические значения случайного члена.

Задача регрессионного анализа состоит в получении оценок β_1и β_2 и, следовательно, в определении положения прямой по точкам *P*.

Почему же существует случайный член? Имеется несколько причин.

1. *Невключение объясняющих переменных.* Соотношение между Y и *X*почти наверняка является очень большим упрощением. В действительности существуют другие факторы, влияющие на Y, которые не учтены в уравнении (1.1), влияние этих факторов приводит к тому, что наблюдаемые точки лежат вне прямой. Часто происходит так, что имеются переменные, которые мы хотели бы включить в регрессионное уравнение, но не можем этого сделать потому, что не знаем, как их измерить.

Например, далее в этой главе мы будем оценивать функции заработка, связывающие часовые заработки с продолжительностью образования. Мы хорошо знаем, что срок обучения является не един-

ственным фактором, влияющим на заработки, и в конечном счете мы усовершенствуем модель, включив в нее и другие переменные, такие как, например, трудовой стаж.

Тем не менее, даже наилучшим образом специфицированная функция заработка объясняет не более половины разброса уровня заработков.

Многие другие факторы влияют на возможность получения хорошей работы, такие как, например, неизмеримые характеристики индивида или даже чистый фактор удачи в смысле нахождения данным индивидом работы, наилучшим образом соответствующей его индивидуальным способностям. Все эти прочие факторы вносят свой вклад в случайный член.

2. Агрегирование переменных. Во многих случаях рассматриваемая зависимость − это попытка объединить вместе некоторое количество микроэкономических соотношений. Например, функция суммарного потребления − это попытка суммарного выражения совокупности решений отдельных индивидуумов о расходах. Так как отдельные соотношения, вероятно, имеют разные параметры, любая попытка определить соотношение между суммарными расходами и совокупным доходом является лишь аппроксимацией. Наблюдаемое расхождение при этом приписывается наличию случайного члена.

3. *Неправильное описание структуры модели.* Структура модели может быть описана неправильно или не вполне правильно. Здесь можно привести один из многих возможных примеров. Если зависимость относится к данным о временном ряде, то значение Y может зависеть не от фактического значения X_i а от значения, которое ожидалось в предыдущем периоде. Если ожидаемое и фактическое значения тесно связаны, то будет казаться, что между Y и X существует зависимость, но это будет лишь аппроксимация, и расхождение вновь будет связано с наличием случайного члена.

4. *Неправильная функциональная спецификация.* Функциональное соотношение между Y и X математически может быть определено неправильно. Например, истинная зависимость может не являться линейной, а быть более сложной. Нелинейные зависимости будут рассмотрены в главе 4. Безусловно, надо постараться избежать возникновения этой проблемы, используя подходящую математическую формулу, но любая самая изощренная формула является лишь приближением, и существующее расхождение вносит вклад в остаточный член.

5. *Ошибки измерения.* Если в измерении одной или более взаимосвязанных переменных имеются ошибки, то наблюдаемые значения не

будут соответствовать точному соотношению, и существующее расхождение будет вносил вклад в остаточный член.

Случайный член является суммарным проявлением всех этих факторов Очевидно, что если бы вас интересовало только измерение влияния X на У, то было бы значительно удобнее, если бы случайного члена не было. Если бы он отсутствовал, то точки P на рис. 1.1 совпадали бы с точками *Q,* и мы бы знали. что любое изменение в Y от наблюдения к наблюдению вызвано изменением *X* и смогли бы точно вычислить β_1и β_2.

Однако в действительности каждое изменение У отчасти вызвано изменением u, и это значительно усложняет жизнь. По этой причине *u* иногда описывается как «шум».

2. Регрессия методом наименьших квадратов

Допустим, что вы имеете четыре наблюдения для *X* и Y, представленные на рис. 1.1, и перед вами поставлена задача получить оценки значений β_1и β_2 в уравнении (1.1). В качестве грубой аппроксимации вы можете сделать это, отложив четыре точки *P* и построив прямую, в наибольшей степени соответствующую этим точкам. Это сделано на рис. 1.2. Отрезок, отсекаемый прямой

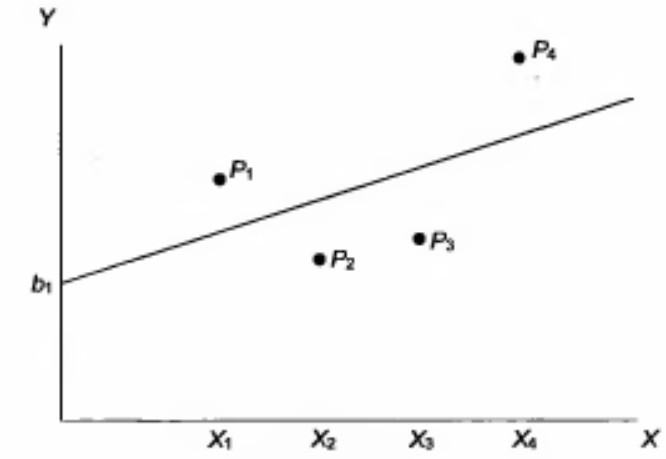

Рисунок 1.2. Оцененная регрессионная прямая

на оси *Y,* представляет собой оценку β_1, он обозначен b_1 а угловой коэффициент прямой представляет собой оценку коэффициента наклона β_2, он обозначен как b_2. Прямая, называемая **оцениваемой моделью,** описывается как

$$\hat{Y}_i = b_1 + b_2 X_i \qquad (1.2)$$

где шляпка над *Y* означает, что это **оцененное значение** *Y* в зависимости от *X*, а не истинное его значение. На рис. 1.3 оцененные, или «теоретические», точки представлены как R_1 - R_4.

С самого начала необходимо признать, что вы никогда не сможете определить истинные значения β_1и β_2 при попытке определить положение искомой прямой. Можно получить только оценки, и они могут быть хорошими или плохими. Иногда ваши оценки могут быть абсолютно точными, но это возможно лишь в результате случайного совпадения, и даже в этом случае у вас не будет способа узнать, что ваши оценки абсолютно точны.

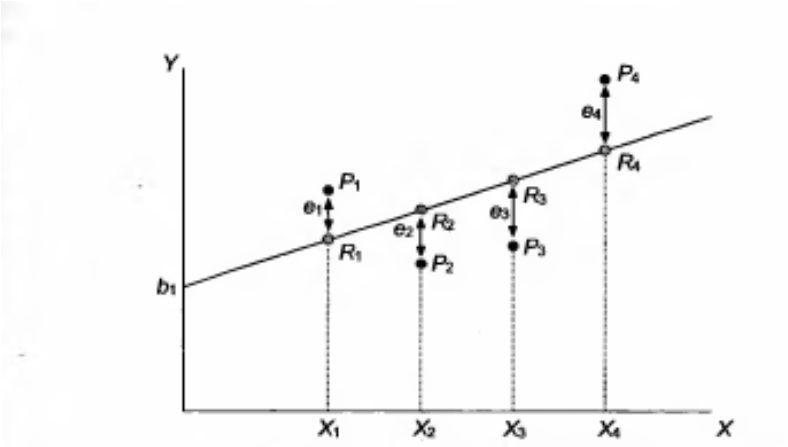

Рисунок 1.3. Оцененная по точкам наблюдений линия регрессии (показаны также остатки)

Это справедливо и при использовании более совершенных методов. Построение линии регрессии на глаз является достаточно субъективным. Более того, как мы увидим в дальнейшем, оно просто невозможно, если переменная Y зависит не от одной, а от двух или более независимых переменных. Возникает вопрос: существует ли способ достаточно точной оценки β_1и β_2 алгебраическим путем?

В качестве первого шага нужно определить, что понимается под **остатком** для каждого наблюдения. Это разность между действительной величиной Y в соответствующем наблюдении и расчетным значением по уравнению регрессии, то есть расстояние по вертикали между P_i и R_i в наблюдении *i*. Оно будет обозначаться как e_i:

$$e_i = Y_i - \hat{Y}_i \qquad (1.3)$$

Остатки для четырех наблюдений показаны на рис. 1.3. Подставив (1.2) в (1.3), мы получим:

$$e_i = Y_i - b_1 - b_2 X_i \qquad (1.4)$$

и, значит, остаток в каждом наблюдении зависит от нашего выбора

значений b_1 и b_2. Очевидно, мы хотим построить линию регрессии, то есть выбрать значения b_1 и b_2таким образом, чтобы эти остатки были минимальными. Очевидно также, что линия, хорошо соответствующая одним наблюдениям, не будет соответствовать другим, и наоборот. Необходимо выбрать какой-то критерий подбора, который будет одновременно учитывать величину всех остатков.

Существует целый ряд возможных критериев, одни из которых работают лучше других. Например, бесполезно минимизировать сумму остатков. Сумма будет автоматически равна нулю, если вы сделаете $b_1 = \acute{Y}$ и $b_2 = 0$, получив горизонтальную линию $Y = \acute{Y}$. В этом случае положительные остатки точно уравновесят отрицательные, но, несмотря на это, данная прямая не соответствует точкам наблюдений.

Один из способов решения поставленной проблемы состоит в минимизации ***RSS (residual sum of squares),*** **суммы квадратов остатков.** Для рис. 1.3 справедлива формула:

$$RSS = e_1^2 + e_2^2 + e_3^2 + e_4^2 \qquad (1.5)$$

В соответствии с этим критерием, чем меньше *RSS,* тем больше соответствие. Если *RSS* равно нулю, то получено абсолютно точное соответствие, так как это означает, что все остатки равны нулю. В этом случае линия будет проходить через все точки, однако, вообще говоря, это невозможно из-за наличия случайного члена.

Существуют и другие достаточно разумные решения, однако при выполнении определенных условий метод **наименьших** квадратов дает несмещенные и эффективные оценки b_1 и b_2. По этой причине метод наименьших квадратов является наиболее популярным при использовании методов регрессионного анализа в относительно простых приложениях. Здесь рассматривается **обычный** метод **наименьших квадратов (МНК,** или ***OLS − ordinary least squares***). В последующих темах будут рассмотрены другие его варианты, которые могут быть использованы для решения некоторых специальных проблем.

Пример 1

Приведем простой пример всего с двумя наблюдениями для того, чтобы продемонстрировать механику процесса: как показано на рис.1.4, наблюдаемое значение Y равно 3, когда X равен 1; и Y= 5 при $X = 2$.

Предположим, что истинная модель имеет вид

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad (1.6)$$

и оценим коэффициенты b_1 и b_2 уравнения

$$\hat{Y}_i = b_1 + b_2 X_i \quad (1.7)$$

Очевидно, что при наличии всего двух наблюдений мы можем получить точное соответствие, проведя линию регрессии через две точки, однако сделаем вид, что мы этого не понимаем. Вместо этого придем к тому же выводу, используя метод рецессии.

Если $X = 1$, то $\acute{Y} = (\beta_1 + \beta_2)$ в соответствии с уравнением регрессии. Если $X=2$, то $\acute{Y} = (\beta_1 + 2\beta_2)$. Следовательно, мы можем сформировать табл. 1.1. Таким образом, остаток для первого наблюдения (e_1), который задается выражением $(Y_1 - \acute{Y}_1)$, равен $(3 - \beta_1 - \beta_2)$, и e_2, заданный выражением $(Y_2 - \acute{Y}_2)$, равен $(5 - \beta_1 - \beta_2)$. Следовательно,

$$RSS = (3 - b_1 - b_2)^2 + (5 - b_1 - 2b_2)^2 = 9 + b_1^2 + b_2^2 - 6b_1 - 6b_2 + 2b_1b_2$$
$$+ 25 + b_1^2 + 4b_2^2 - 10b_1 - 20b_2 + 4b_1b_2 = 34 + 2b_1^2 + 5b_2^2 - 16b_1 -$$
$$26b_2 + 6b_1b_2 \quad (1.8)$$

Теперь мы хотим выбрать такие значения b_1 и b_2 чтобы значение *RSS было* минимальным. Для этого мы используем дифференциальное исчисление и находим значения b_1 и b_2 удовлетворяющие следующим соотношениям:

$$\frac{\partial RSS}{\partial b_1} = 0 \text{ и } \frac{\partial RSS}{\partial b_2} = 0 \quad (1.9)$$

Взяв частные производные, получаем

$$\frac{\partial RSS}{\partial b_1} = 4b_1 + 6b_2 - 16 \quad (1.10)$$

и

$$\frac{\partial RSS}{\partial b_2} = 10b_2 + 6b_1 - 26 \quad (1.11)$$

Таким образом, мы имеем

$$2b_1 + 3b_2 - 8 = 0 \quad (1.12)$$

и

$$3b_1 + 5b_2 - 13 = 0 \quad (1.13)$$

Решив эти два уравнения, получим $b_1 = 1$ и $b_2 = 2$, следовательно, уравнение регрессии будет иметь следующий вид:

$$\hat{Y}_i = 1 + 2X_i \quad (1.14)$$

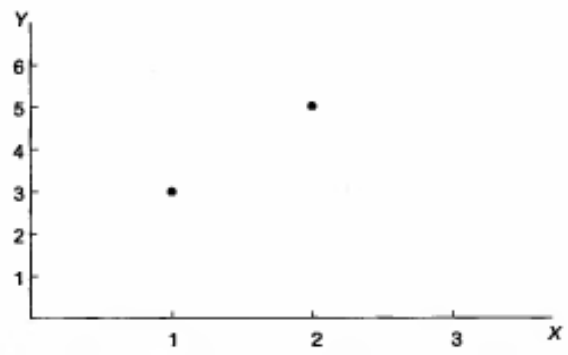

Рисунок 1.4. Пример с двумя наблюдениями

Таблица 1.1. Пример с двумя наблюдениями

X	Y	$\hat{Y}$	e
1	3	$b_1 + b_2$	$3 - b_1 - b_2$
2	5	$b_1 + 2b_2$	$5 - b_1 - 2b_2$

Для того чтобы проверить, что мы пришли к правильному выводу, вычислим остатки:

$$e_1 = 3 - b_1 - b_2 = 3 - 1 - 2 = 0 \qquad (1.15)$$

$$e_2 = 5 - b_1 - 2b_2 = 5 - 1 - 4 = 0 \qquad (1.16)$$

Таким образом, оба остатка равны нулю, что означает, что линия проходит точно через обе точки. И это мы, разумеется, знали с самого начала.

Пример 2

Используем пример, рассмотренный выше, и добавим третье наблюдение: Y равен 6 при X= 3. Три наблюдения, показанные на рис.1.5, не лежат на одной прямой, поэтому точное соответствие получить невозможно.

В этом случае для вычисления положения прямой мы используем регрессию по методу наименьших квадратов. Начнем с задания стандартного уравнения

$$\hat{Y}_i = b_1 + b_2 X_i \qquad (1.17)$$

Для значений X, равных 1, 2 и 3, расчетные значения Y равны соответственно $(b_1 + b_2)$, $(b_1 + 2b_2)$ и (b_1+3b_2), они приведены в табл. 1.2.

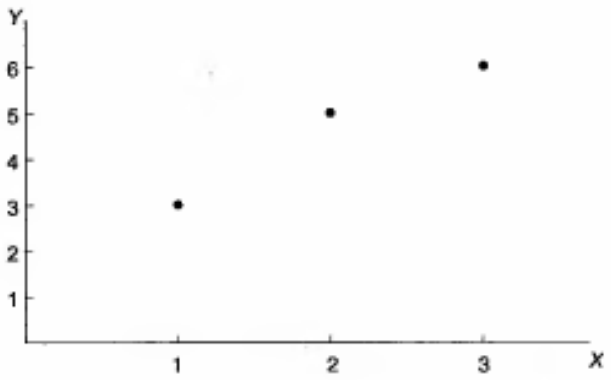

Рисунок 1.5. Пример с тремя наблюдениями

Таблица 1.2. Пример с тремя наблюдениями

X	Y	$\hat{Y}$	e
1	3	$b_1 + b_2$	$3 - b_1 - b_2$
2	5	$b_1 + 2b_2$	$5 - b_1 - 2b_2$
3	6	$b_1 + 3b_2$	$6 - b_1 - 3b_2$

Следовательно,

$$RSS = (3 - b_1 - b_2)^2 + (5 - b_1 - 2b_2)^2 = 9 + b_1^2 + b_2^2 - 6b_1 - 6b_2 + 2b_1b_2 + 25 + b_1^2 + 4b_2^2 - 10b_1 - 10b_2 + 4b_1b_2 + 36 + b_1^2 + 9b_2^2 - 2b_1 - 36b_2 + 6b_1b_2 = 70 + 3b_1^2 + 14b_2^2 - 28b_1 - 62b_2 + 12b_1b_2 \quad (1.18)$$

Условия первого $\frac{\partial RSS}{\partial b_1} = 0$ и $\frac{\partial RSS}{\partial b_2} = 0$ дают

$$6b_1 + 12b_2 - 28 = 0 \quad (1.19)$$
$$12b_1 + 28b_2 - 62 = 0 \quad (1.20)$$

Решая систему этих двух уравнений, получим $b_1 = 1{,}67$ и $b_2 = 1{,}50$. Следовательно, уравнение регрессии имеет следующий вид:

$$\hat{Y}_i = 1{,}67 + 1{,}50X_i \quad (1.21)$$

Три наблюдения и линия регрессии представлены на рис. 1.6.

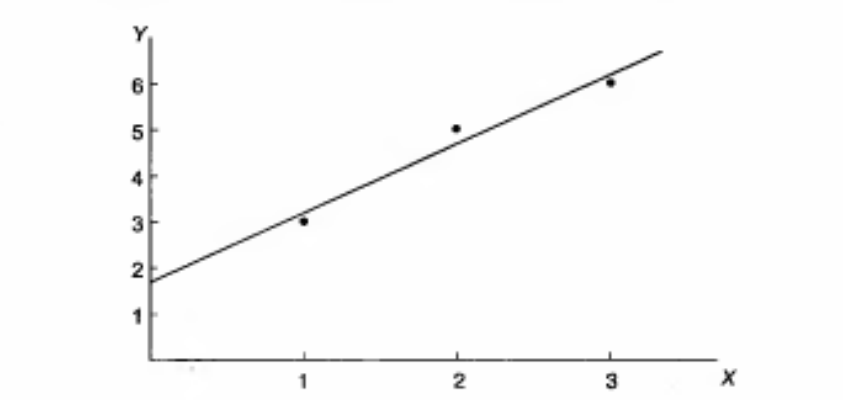

Рисунок 1.6. Пример построения линии регрессии с тремя наблюдениями

Два разложения для зависимой переменной

Выше были рассмотрены два способа разложения величины зависимой переменной в регрессионной модели.

Они будут использоваться и далее, и поэтому важно их правильное понимание и концептуальное разграничение.

Первое из разложений связано с процессом, в соответствии с которым генерируются величины *Y:*

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad (1.40)$$

В наблюдении / величина Y_i генерируется как сумма двух компонентов: нестохастического компонента $\beta_1 + \beta_2 X_i$. и случайного члена u_i. Это разложение — чисто теоретическое.

Мы будем использовать его при анализе свойств оценок регрессии. Оно проиллюстрировано на рис. 1.7а, где *QT*— нестохастическая составляющая Y и *PQ* — случайный член.

Другое разложение относится к линии регрессии:

$$Y_i = \hat{Y}_i + e_i = b_1 + b_2 X_i + e_i \qquad (1.41)$$

Как только мы выбрали значения b_1 и b_2, каждая величина Y разлагается на расчетное (теоретическое) значение $\hat{Y}_i$ и остаток e_i. Это разложение практически выполнимо, но оно в определенной степени произвольно, поскольку зависит от нашего критерия для определения b_1 и b_2, и на него неизбежно будут влиять конкретные значения случайного члена в наблюдениях выборки. Это показано на рис. 1.7б, где *RT*— расчетное значение *и PR* — остаток.

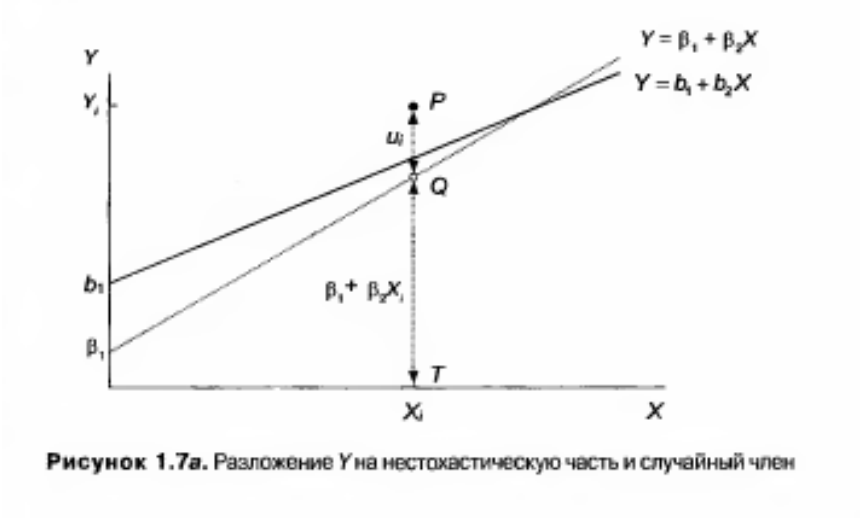

Рисунок 1.7а. Разложение *Y* на нестохастическую часть и случайный член

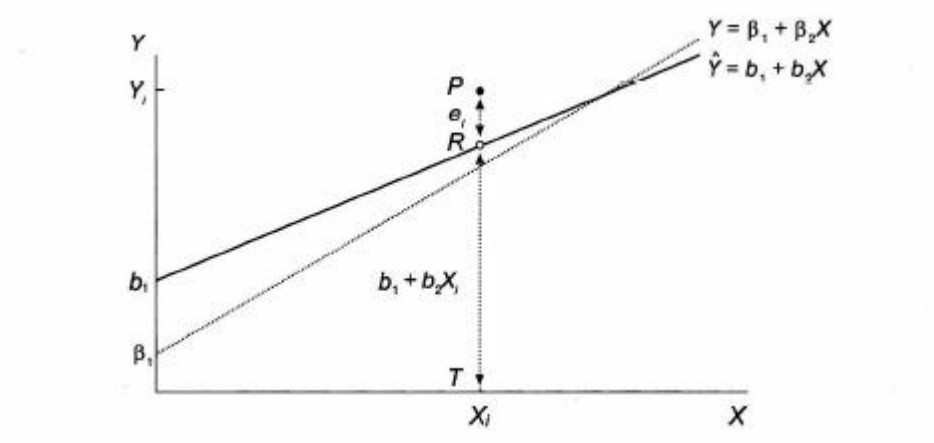

Рисунок 1.7б. Разложение У на расчетное значение и остаток

3. Интерпретация уравнения регрессии

Существуют два этапа интерпретации уравнения регрессии.

Первый этап состоит в словесном истолковании уравнения так, чтобы это было понятно человеку, не являющемуся специалистом в области эконометрики.

На втором этапе необходимо решить, следует ли ограничиться этим или следует провести более детальное исследование зависимости. Оба этапа важны. Второй этап мы рассмотрим несколько позже, а пока обратим основное внимание на первый этап. Это будет проиллюстрировано на примере функции заработка, часового заработка в 2002 г. *(EARNINGS),* измеренного в долларах США, для которой строится регрессионная зависимость от продолжительности обучения *S,* опре-

деляемой как число завершенных лет обучения для 540 респондентов из Национального опроса молодежи в США *(NLSY)* в 1979 г. Эта база данных используется во многих других примерах и упражнениях книги. В Приложении *B* содержится ее описание. Приведенная ниже регрессия использует набор данных 21 из базы данных *EAEF (см. файл eaef21.dta на практических занятиях).* В табл. 1.3 приведена распечатка результатов оценивания данной регрессии с помощью программы Stata. Соответствующая диаграмма рассеяния и линия регрессии показаны на рис. 1.8.

На данном этапе игнорируйте все, кроме столбца с заголовком «coef.» в нижней половине таблицы. В нем показаны оценки коэффициента при переменной *S* и свободного члена, и, таким образом, имеем следующее оцененное уравнение:

$$EARNINGS = -13{,}93 + 2{,}46\,S \qquad (1.42)$$

Интерпретируя оцененное уравнение буквально, можно сказать, что коэффициент наклона показывает, что при увеличении *S* на одну единицу (измерения *S) EARNINGS* возрастает на 2,46 единиц (измерения *EARNINGS).* Поскольку *S* измеряется в годах, а *EARNINGS* измеряется в долларах в час, коэффициент при *S* показывает, что часовые заработки возрастают на 2,46 долл. на каждый дополнительный год учебы.

```
. regress  EARNINGS S
```

Source	SS	df	MS
Model	19321.5589	1	19321.5589
Residual	92688.6722	538	172.283777
Total	112010.231	539	207.811189

Number of obs	540
F(1, 538)	112.15
Prob > F	0.0000
R-squared	0.1725
Adj R-squared	0.1710
Root MSE	13.126

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	2.455321	.2318512	10.59	0.000	1.999876	2.910765
_cons	-13.93347	3.219851	-4.33	0.000	-20.25849	-7.608444

Что можно сказать о постоянном члене в уравнении? Формально говоря, он показывает прогнозируемый уровень *EARNINGS* при значении *HGC,* равном нулю. Иногда постоянный член имеет ясный смысл, иногда — нет. Если значения объясняющих переменных в выборке находятся достаточно далеко от нуля, то экстраполирование линии регрессии назад до нуля может породить проблемы. Даже если линия регрессии дает хорошее соответствие для наблюдаемой выбор-

ки, нет гарантии, что так же будет при экстраполяции влево или вправо.

В данном случае буквальная интерпретация постоянного члена привела бы к бессмысленному выводу о том, что индивид с нулевым образованием имел бы часовой заработок в размере 13,93 долл. В нашем наборе данных никто из респондентов не имеет менее семи лет образования, поэтому неудивительно, что экстраполяция до нуля привела к проблемам.

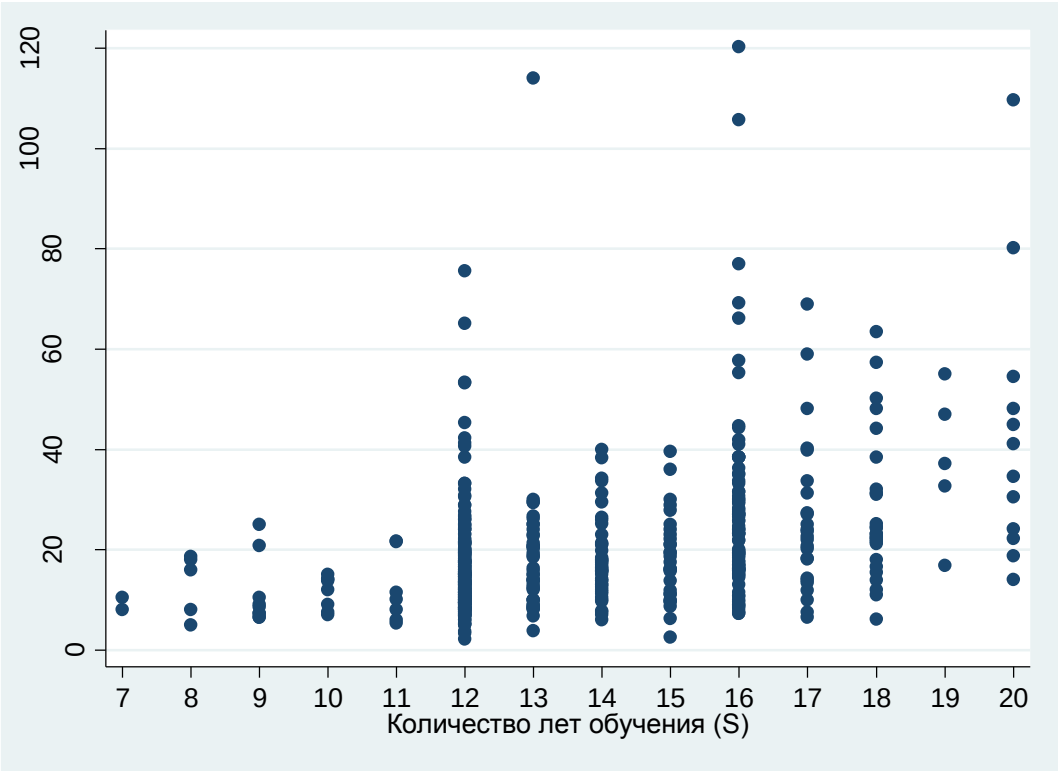

Рисунок 1.8. Простейшая функция заработка

Вставка 1.1 дает общее руководство по интерпретации уравнения регрессии, когда переменные измерены в естественных единицах.

Вставка 1.1. ***Интерпретация линейного уравнения регрессии***

Представим простой способ интерпретации коэффициентов линейного уравнения регрессии:

$$\hat{Y}_i = b_i + b_2 x_i$$

если Y и X — переменные с простыми, естественными единицами измерения.

Во-первых, мы можем сказать, что увеличение X на одну единицу (в единицах измерения переменной *X)* приведет к увеличению значения *Y на* b_2 единиц (в единицах измерения переменной *Y).* Вторым шагом является проверка, каковы действительно единицы измерения *X* и *Y,* и замена слова «единица» фактической единицей измерения. Третьим шагом является проверка возможности более удобного представления результата без потери его сущности.

Постоянная b_1 дает прогнозируемое значение *У (в* единицах *У),* если *X* равен нулю. Это может иметь или не иметь ясного смысла в зависимости от контекста.

При интерпретации уравнения регрессии важно помнить о трех вещах. Во-первых, поскольку *b1* есть всего лишь оценка b_i а *b2* — оценка b_2, интерпретация уравнения — также всего лишь оценка. Во-вторых, уравнение регрессии описывает лишь общую тенденцию в выборке. Каждый частный случай находится под влиянием случайных факторов. В-третьих, интерпретация основана на предположении, что уравнение правильно специфицировано. В действительности такая интерпретация функции заработка довольно наивна. Мы несколько раз вновь рассмотрим ее в последующих главах. Вам потребуется провести аналогичные эксперименты с использованием одного из других наборов данных *EAEF*, описанных в Приложении *B.*

Оценив регрессию, естественно задать вопрос о том, насколько аккуратными оказались наши оценки. Этот важный вопрос будет обсужден в следующей главе.

4. Качество оценивания: коэффициент R^2

Цель регрессионного анализа состоит в объяснении поведения зависимой переменной ***Y.*** В любой данной выборке значение ***Y*** оказывается сравнительно низким в одних наблюдениях и сравнительно высоким — в других. Мы хотим знать, почему это так. Разброс значений ***Y*** в любой выборке можно суммарно описать с помощью $\sum(Y_i - \bar{Y})^2$, суммы квадратов отклонений от выборочного среднего. Мы должны уметь рассчитывать величину и структуру этой статистики.

Выше было показано, что мы можем разбить значение ***Y*** **в** каждом наблюдении на две составляющие - $\widehat{Y_i}$ и e_i:

$$Y_i = \hat{Y}_i + e_i \qquad (1.43)$$

Это соотношение можно использовать для разложения $\sum(Y_i - \bar{Y})^2$:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}((\hat{Y}_i + \bar{e}))^2 = \sum_{i=1}^{n}((\hat{Y}_i - \bar{Y}) + \mathrm{e_i})^2) \qquad (1.44)$$

На втором шаге мы использовали тот факт, что $\bar{e}$= 0 и $\bar{Y}$ = **Y,** продемонстрированный во вставке 1.2. Следовательно $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}((\hat{Y}_i - \bar{Y}))^2 + \sum_{i=1}^{n}\mathrm{e_i}^2 + 2\sum_{i=1}^{n}((Y_i - \bar{Y})\mathrm{e_i}) =$
$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + \sum_{i=1}^{n}\mathrm{e_i}^2 + 2\sum_{i=1}^{n}\hat{Y}_i + e_i - 2\bar{Y}\sum_{i=1}^{n}e_i$

Как показано во Вставке 1.2, $\sum \boldsymbol{Y_i e_i}$= 0 и $\sum \boldsymbol{e_i}$ = $\overline{\text{пё}}$= 0. Следовательно,

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + \sum_{i=1}^{n}e_i^2$$

Таким образом, имеем следующее разложение:

TSS=ESS+RSS

где ***TSS*, общая сумма квадратов,** дана в левой части уравнения, а ***ESS*, «объясненная» сумма квадратов,** и ***RSS*, остаточная («необъясненная») сумма квадратов,** представляют два слагаемых в его правой части. ***(Замечание:*** слова «объясненная» и «необъясненная» заключены в кавычки, поскольку объяснение может оказаться мнимым. Величина сможет в действительности зависеть от некоторой другой переменной ***Z,*** а переменная ***X*** может служить в качестве замещающей переменной для Z (позже это будет пояснено более подробно). Было бы более правильно использовать выражение «видимое объяснение» вместо «объяснение».)

Согласно (1.46), $\sum(\hat{Y}_i - \bar{Y})^2 / (Y_i - \bar{Y})^2$ это часть общей суммы квадратов, объясненной уравнением регрессии. Это отношение известно как **коэффициент детерминации,** и его обычно обозначают как R^2**:**

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{n}\sum(\hat{Y}_i - \bar{Y})^2}{\sum(\hat{Y}_i - \bar{Y})^2}$$

Вставка 1.2. Четыре полезных результата относительно регрессий, оцениваемых по обычному МНК:

(1)$\bar{\mathrm{e}}$***=0, (2)*** $\bar{\hat{\mathrm{Y}}} = \bar{\mathrm{Y}}$***,(3)*** $\sum_{i=1}^{n}x_i e_i = 0$, *(4)* $\sum_{i=1}^{n}\hat{Y}_i e_i = 0$

Доказательство (1)

$\mathrm{e_i} = \mathrm{Y_i} - \hat{Y}_i = \mathrm{Y_i} - b_1 - b_2\mathrm{X}_2$

Откуда

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} Y_i - n\,b_1 - \sum_{i=1}^{n} X_i$$

Разделив на n, получаем

$$\bar{e} = \bar{Y}_i - b_1 - b_2 X_2 = \bar{Y} - (Y - b_2\bar{X}) - b_2\bar{X} = 0$$

Доказательство (2)

$$e_i = \bar{Y} - \bar{\hat{Y}}_i$$

Однако $\bar{e}$***=0, откуда*** $\bar{\hat{Y}}_i$=$\bar{Y}$

Доказательство (3)

$\sum_{i=1}^{n} x_i e_i = \sum_{i=1}^{n} x_i(Y_i - b_1 - b_2 X_i)$=$\sum_{i=1}^{n} X_i Y_i$-$b_i \sum_{i=1}^{n} X_i$-
$b_2 \sum_{i=1}^{n} X^2{}_i$=*0*

Финальный шаг использует уравнение (1.29.)

Доказательство (4)

$$\sum_{i=1}^{n} \hat{Y}_i e_i = \sum_{i=1}^{n} e_i(b_1 + b_2 X_i) = b_1 \sum_{i=1}^{n} e_i + b_2 \sum_{i=1}^{n} X_i e_i = 0$$

$\sum e_i$=0, так как $\bar{e} = 0$ и $\sum e_i X_i$=**0 из 3.**

Распечатка регрессионного анализа всегда включает $\boldsymbol{R^2}$и может также содержать лежащий в его основе анализ дисперсии. Таблица 1.4 воспроизводит распечатку программы Stata для оценивания функции заработка, приведенной в табл. 1.3.

Таблица 1.4

```
. regress  EARNINGS S
```

Source	SS	df	MS		
Model	19321.5589	1	19321.5589	Number of obs =	540
Residual	92688.6722	538	172.283777	F(1, 538) =	112.15
				Prob > F =	0.0000
				R-squared =	0.1725
				Adj R-squared =	0.1710
Total	112010.231	539	207.811189	Root MSE =	13.126

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	2.455321	.2318512	10.59	0.000	1.999876	2.910765
_cons	-13.93347	3.219851	-4.33	0.000	-20.25849	-7.608444

Колонка с заголовком ***«SS»,*** содержит суммы квадратов. Величина ***ESS,*** описываемая здесь как «моделируемая» (Model) сумма квадратов, равна 19322. Величина ***TSS***(Total) равна 112010. Разделив ***ESS на TSS,*** получим, что $\boldsymbol{R^2}$ = 19322/112010 = 0,1725, что совпадает со значением $\boldsymbol{R^2}$, приведенным в правом верхнем углу таблицы.

Низкое значение $\boldsymbol{R^2}$ частично объясняется тем фактом, что важные переменные, такие как опыт работы, не были учтены в модели. Оно также частично объясняется тем, что ненаблюдаемые характеристики оказывают большое влияние на заработок: R^2 редко бывает выше 0,5, даже когда модель имеет хорошую спецификацию.

Максимально возможное значение R^2 равно единице. Это происходит в том случае, когда линия регрессии точно соответствует всем наблюдениям, так что Y_f — Y_i для всех наблюдений и все остатки равны нулю. Тогда

$\sum(\hat{Y}_i - \bar{Y})^2 = \sum(\mathrm{Y_i} - \bar{Y})^2$, $\sum {e_i}^2$=0 и управление является идеальным. Если в выборке отсутствует видимая связь между Y и X , то R^2 будет близок к нулю.

При прочих равных условиях желательно, чтобы R^2 был как можно больше. В частности, мы заинтересованы в таком выборе коэффициентов $b1$ и $b2$, чтобы максимизировать R^2. Не противоречит ли это нашему критерию, в соответствии с которым $b1$ и $b2$ должны быть выбраны таким образом, чтобы минимизировать сумму квадратов остатков? Нет, легко показать, что эти критерии эквивалентны. На основе (1.46) мы можем записать R^2 как

$$\mathrm{R^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(\mathrm{Y_i} - \bar{Y})}$$

и, таким образом, те значения $b1$ и b_2 которые минимизируют $\sum {e_i}^2$, автоматически максимизируют R^2.

Заметим, что четыре полезных результата во Вставке 1.2 зависят от того, включается ли в модель постоянный член (см. упражнение 1.17). Если его нет, то разложение (1.46) неверно и два определения R^2 в уравнениях (1.48) и (1.49) не эквивалентны. Любое определение R^2 в этом случае может быть обманчивым, и к нему следует относиться с особой осторожностью.

Пример вычисления $\boldsymbol{R^2}$

Вычисление R^2 выполняется на компьютере в рамках программы оценивания регрессии, поэтому данный пример приведен лишь в целях иллюстрации. Будем использовать простейший пример с тремя наблюдениями, описанный выше, где уравнение регрессии

$$\hat{Y}_i = 1{,}6667 + 1{,}5000\mathrm{X_i}$$

построено по наблюдениям *X* и *Y*, приведенным в табл. 1.5. В таблице также даны $\hat{Y}_i$ и e_i для каждого наблюдения. $\sum(\mathrm{Y_i} - \bar{Y})^2$ =4,6667, $\sum(\hat{Y}_i - \bar{Y})^2$=4,5000 и $\sum {e_i}^2$=0,1667. На основании этих значений мы можем вычислить R^2, используя (1.48) или (1.49):

Таблица 1.5.

Анализ дисперсии в примере с тремя наблюдениями

Наблюдение	X	Y	$\hat{Y}$	E	$\mathrm{Y_i} - \bar{Y}$	$\hat{Y}_i - \bar{Y}$	$(\mathrm{Y_i} - \bar{Y})^2$	$(\hat{Y}_i - \bar{Y})^2$	e^2
1	1	3	3,1667	-0,1667	-1,6667	-1,5	2,7778	2,25	0,0278
2	2	5	4,6667	0,3333	0,3333	0,0	0,1111	0,00	0,1111
3	3	6	6,1667	-0,1667	1,3333	1,5	1,7778	2,25	0,0278
Всего	6	14	14				4,6667	4,5	0,1667
Среднее	2	4,6667	4,6667						

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{n}\sum(\hat{Y}_i - \bar{Y})^2}{\sum(\hat{Y}_i - \bar{Y})^2} = \frac{4{,}500}{4{,}6667} = 0{,}96$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(\mathrm{Y_i} - \bar{Y})} = 1 - \frac{0{,}1667}{4{,}667} = 0{,}96$$

Практика в STATA

1. Используйте файл WAGE1.dta для этого упражнения.

(I) Найдите средний уровень образования в выборке. Найдите самое короткое и самое продолжительное обучение.

(II) Найдите среднюю часовую заработную плату в выборке. Кажется ли она Вам высокой или низкой?

(III) Данные о заработной плате представлены в ценах 1976 г. Получите и введите в файл индекс потребительских цен (ИПЦ) за период с 1976 по 2003 год.

(IV) Используйте значения ИПЦ из (III), чтобы найти среднюю часовую заработную плату в 2003 году. Кажется ли Вам разумной сейчас средняя часовая заработная плата?

(V) Сколько женщин в выборке? Сколько мужчин?