

Лекция по эконометрике № 2

Линейная регрессионная модель для случая одной объясняющей переменной

Демидова

Ольга Анатольевна

https://www.hse.ru/staff/demidova_olga

E-mail:demidova@hse.ru

09.09.2019

План лекции № 2

- **Возникновение термина «регрессия»**
- **Теоретическая и выборочная регрессии**
- **Линейность регрессии по переменным и параметрам.**
- **Задача оценивания параметров**
- **Метод наименьших квадратов (МНК)**
- **Система нормальных уравнений и ее решение**
- **Интерпретация оценок МНК на примере**
- **Дисперсионный анализ**
- **R^2**

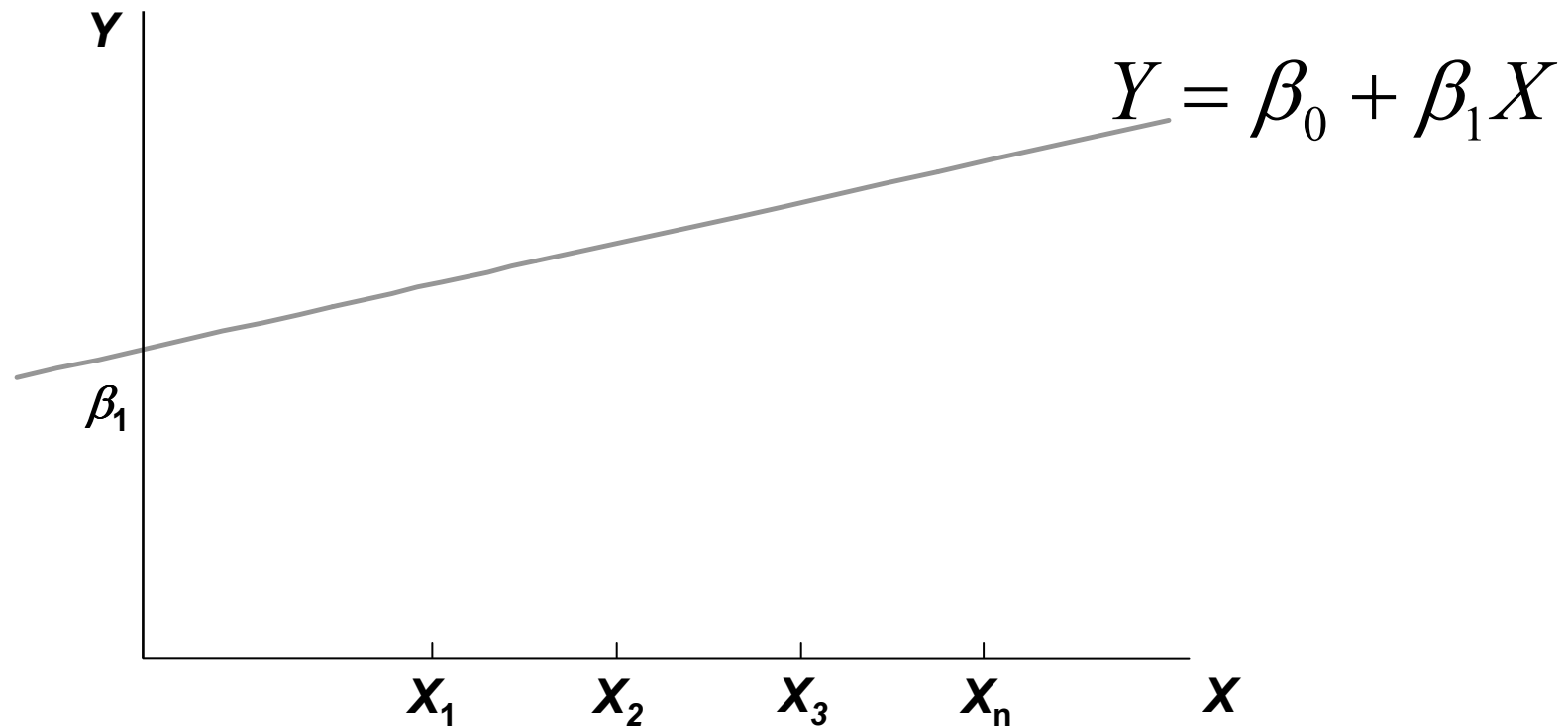
Dictionary definition: “Regression is a backward movement, a retreat, a return to an earlier stage of development”.

Sir Francis Galton (1822-1911) ввел термин «регрессия», изучая зависимость роста детей от роста родителей.

“A regression of children’s height towards the average”.

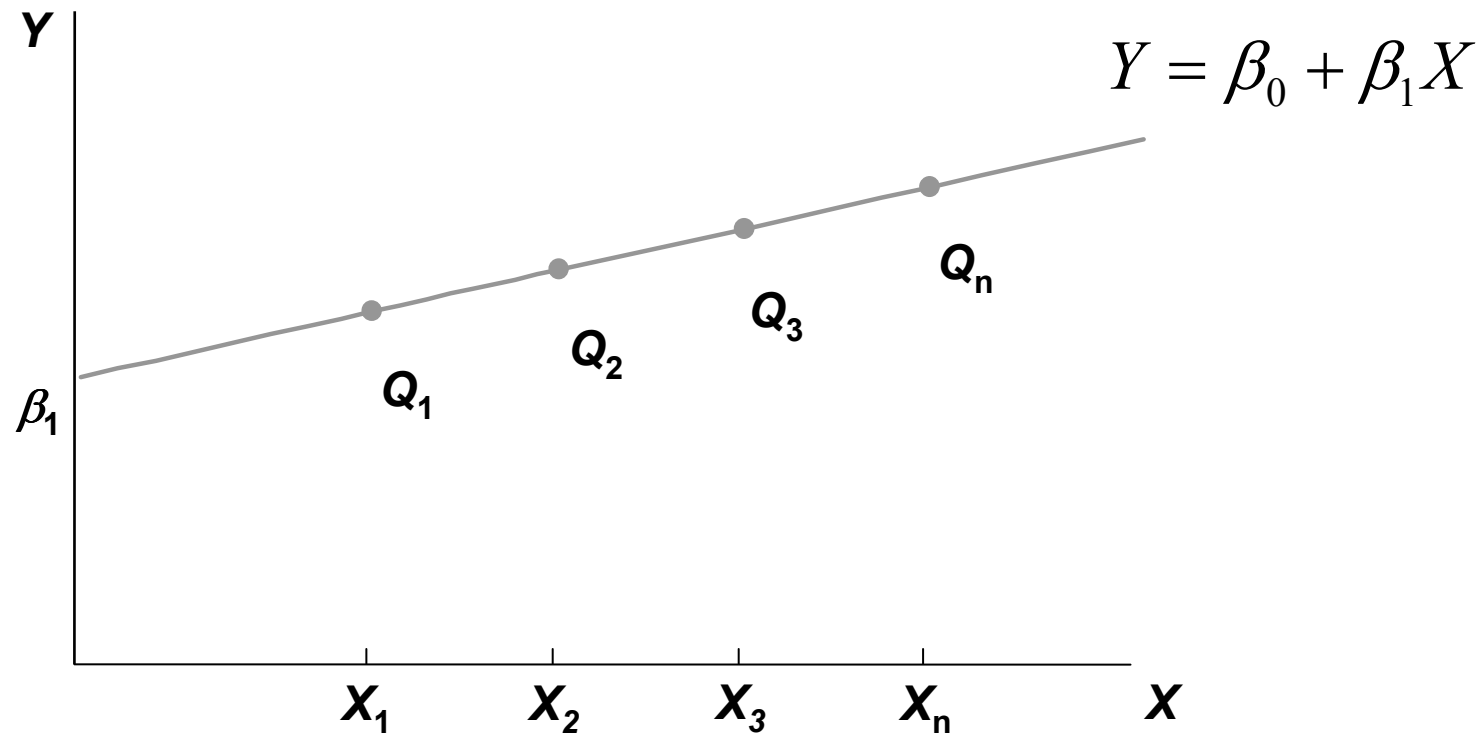
- Линейный регрессионный анализ объединяет широкий круг задач, связанных с построением зависимостей между двумя переменными: X и Y .
- X – независимая, объясняющая, экзогенная переменная, регрессор
- Y – зависимая, объясняемая, эндогенная переменная, regressand.
- На практике исследователь работает с данными (X_i, Y_i) , $i = 1, \dots, n$.

Модель парной регрессии



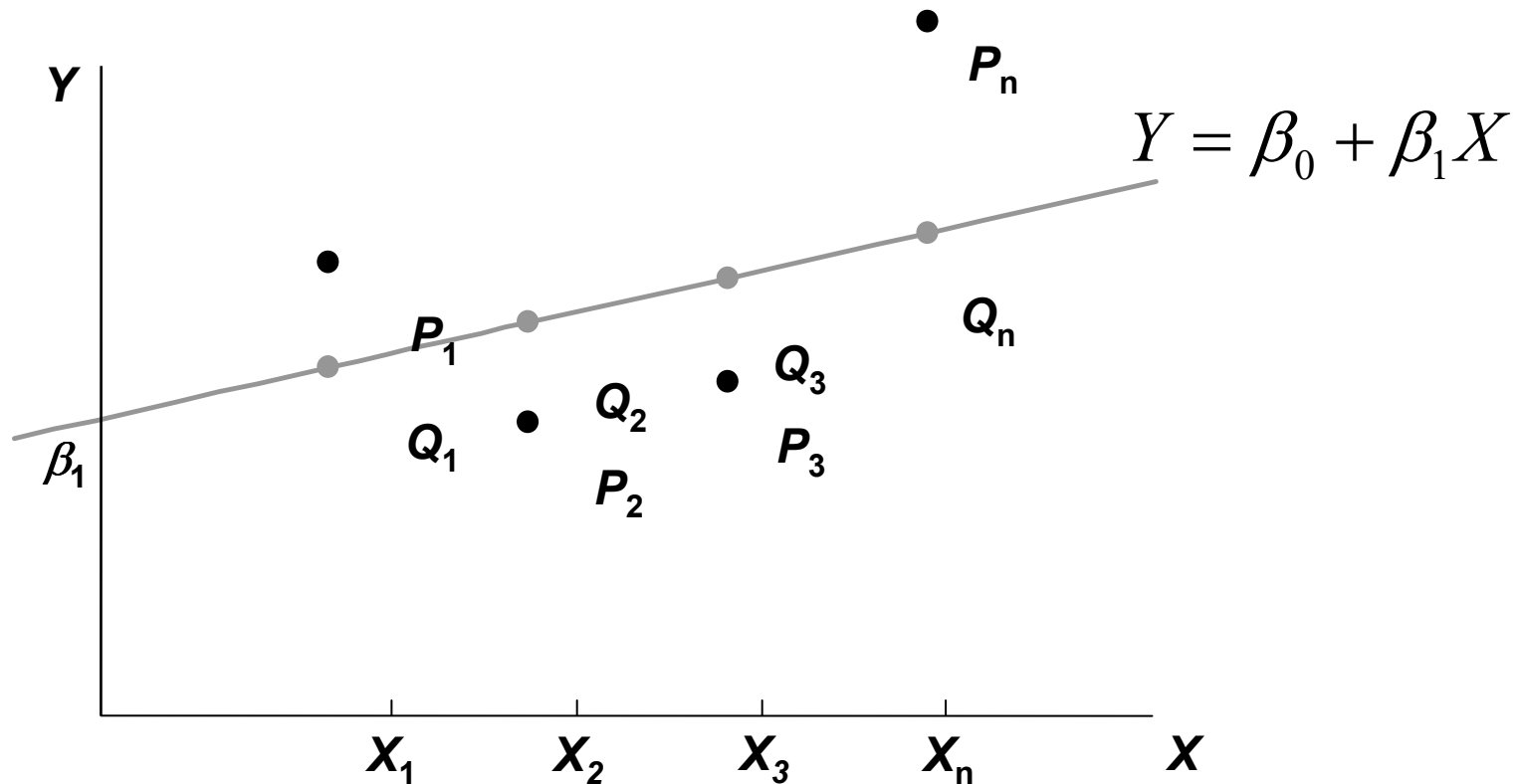
Предположим, что переменная Y является линейной функцией переменной X с неизвестными параметрами β_0 и β_1 , которые мы хотим оценить.

Модель парной регрессии



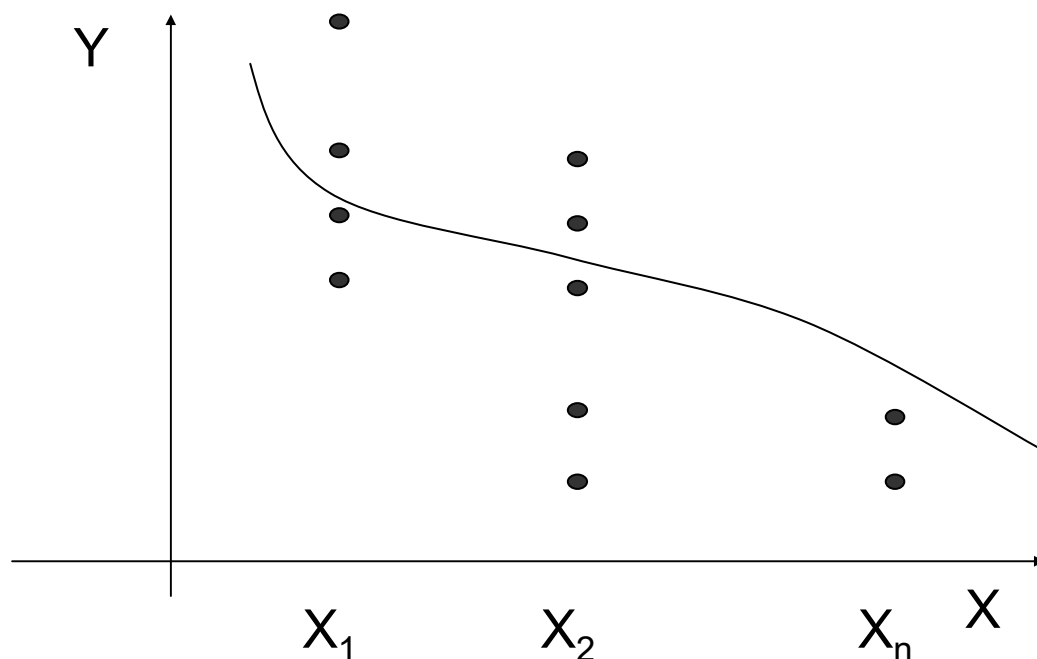
Если все точки расположены на одной прямой линии, то достаточно легко найти оценки параметров β_0 и β_1 .

Модель парной регрессии



Но на практике взаимоотношения между переменными X и Y не являются в точности линейными, т.е. пары наблюдений не располагаются на прямой линии.

Модель парной регрессии



Уравнение теоретической регрессии $Y_i = f(X_i) + \varepsilon_i$, $i = 1, \dots, n$

т.к. при одном и том же X , Y могут быть разные.

Из случайной величины Y выделяем некоторую часть, которая детерминирована x сом,

ε_i - случайная составляющая, добавка.

Модель парной регрессии

$$f(X_i) = E(Y|X = X_i), i = 1, \dots, n$$

$Y_i = E(Y|X = X_i) + \varepsilon_i, i = 1, \dots, n$ – уравнение теоретической регрессии

Какова причина появления случайной составляющей ε_i (возмущения)?

- В модели участвуют не все переменные, влияющие на поведение Y .
- Врожденная неопределенность поведения экономических агентов.
- Мы используем те величины, которые можем измерить, а не те, которые хотелось бы.
- Ошибки измерения.

Модель парной регрессии

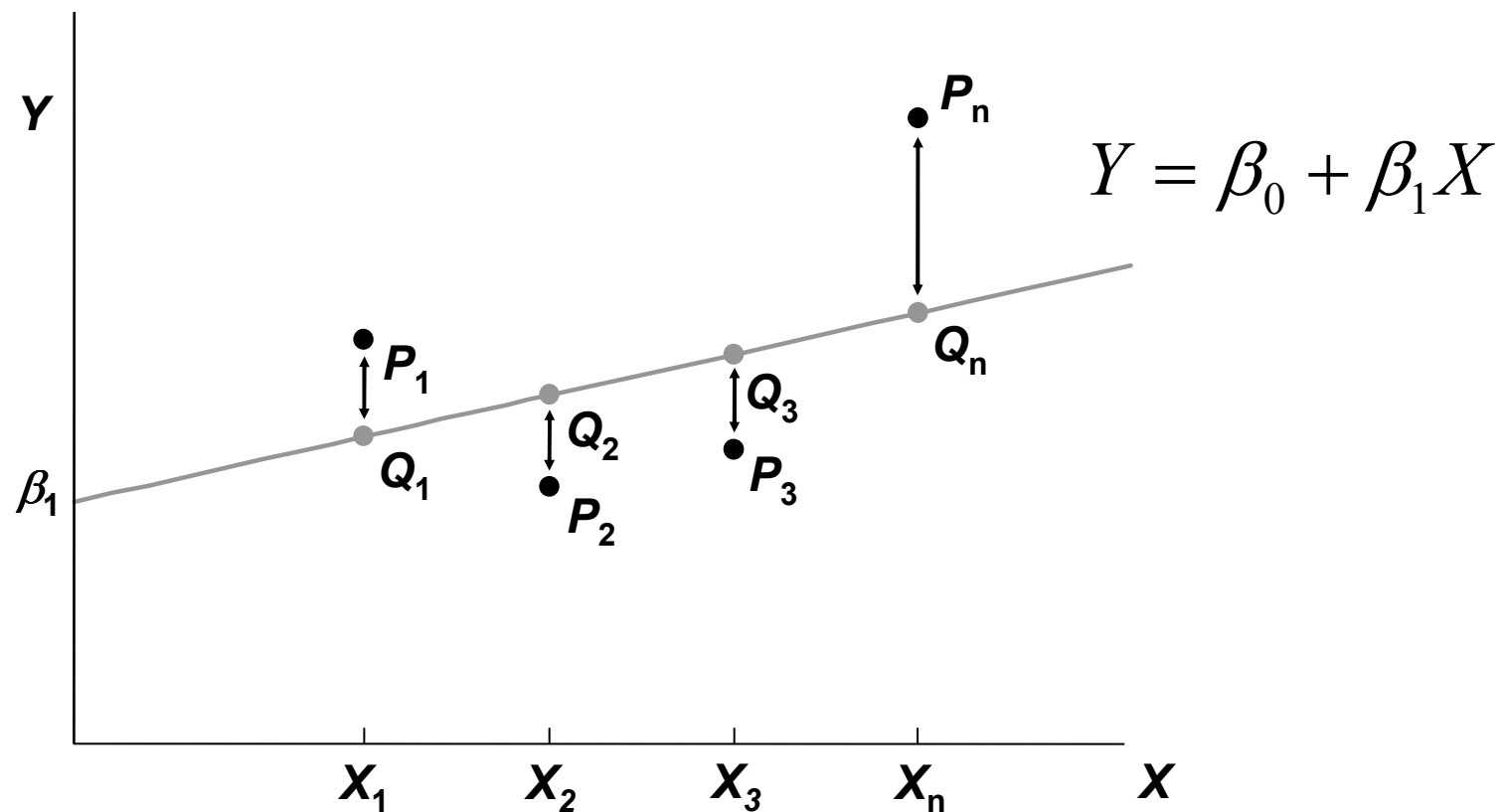
Уравнение теоретической регрессии

$$Y_i = f(X_i) + \varepsilon_i$$

в зависимости от $f(X_i)$ может быть линейным, квадратичным, логарифмическим и т.д.

Рассмотрим линейный случай: $f(x) = \beta_0 + \beta_1 X$ – линейно по X и по параметрам.

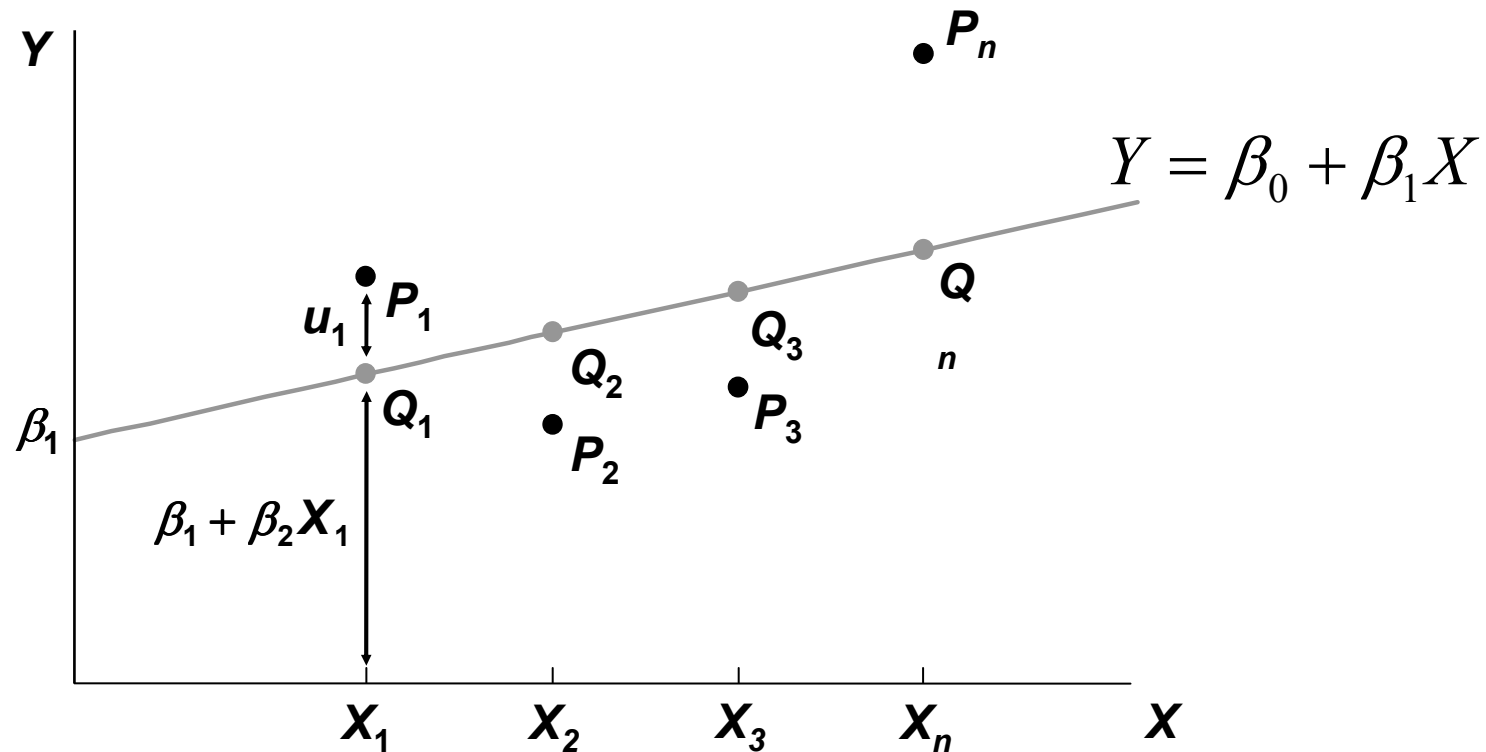
Модель парной регрессии



Специфицируем модель следующим образом:

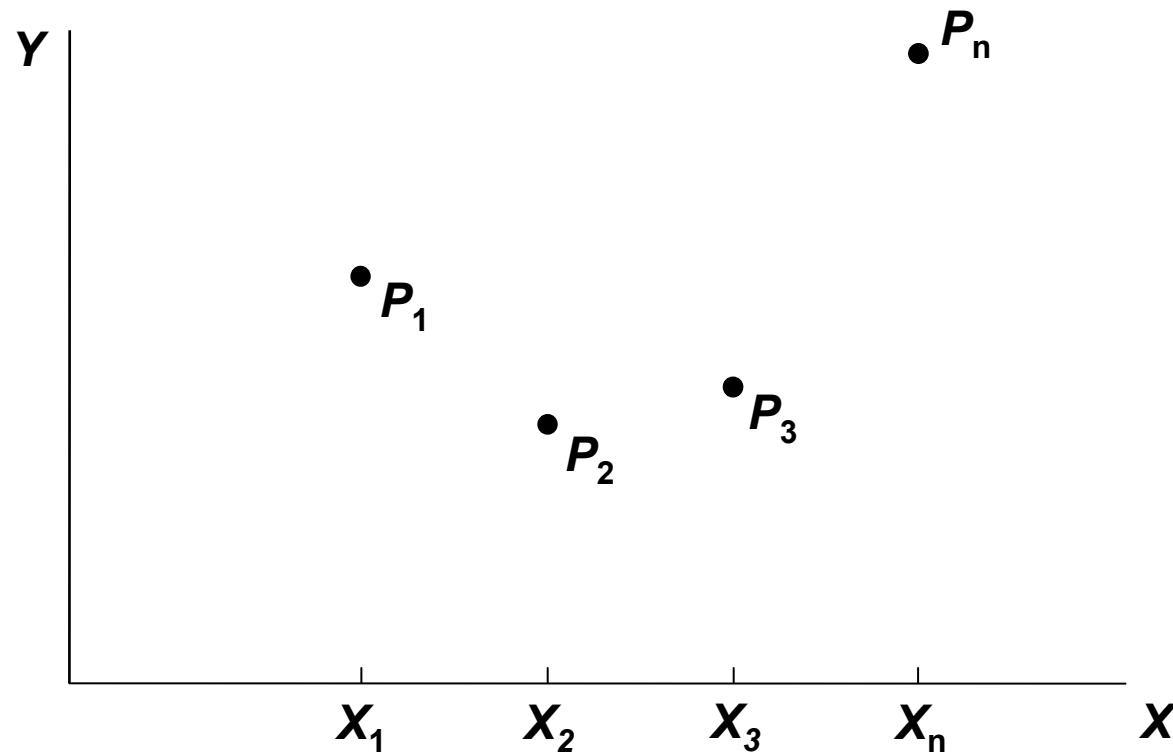
$$Y = \beta_0 + \beta_1 X + \varepsilon, \text{ где } \varepsilon - \text{возмущение.}$$

Модель парной регрессии



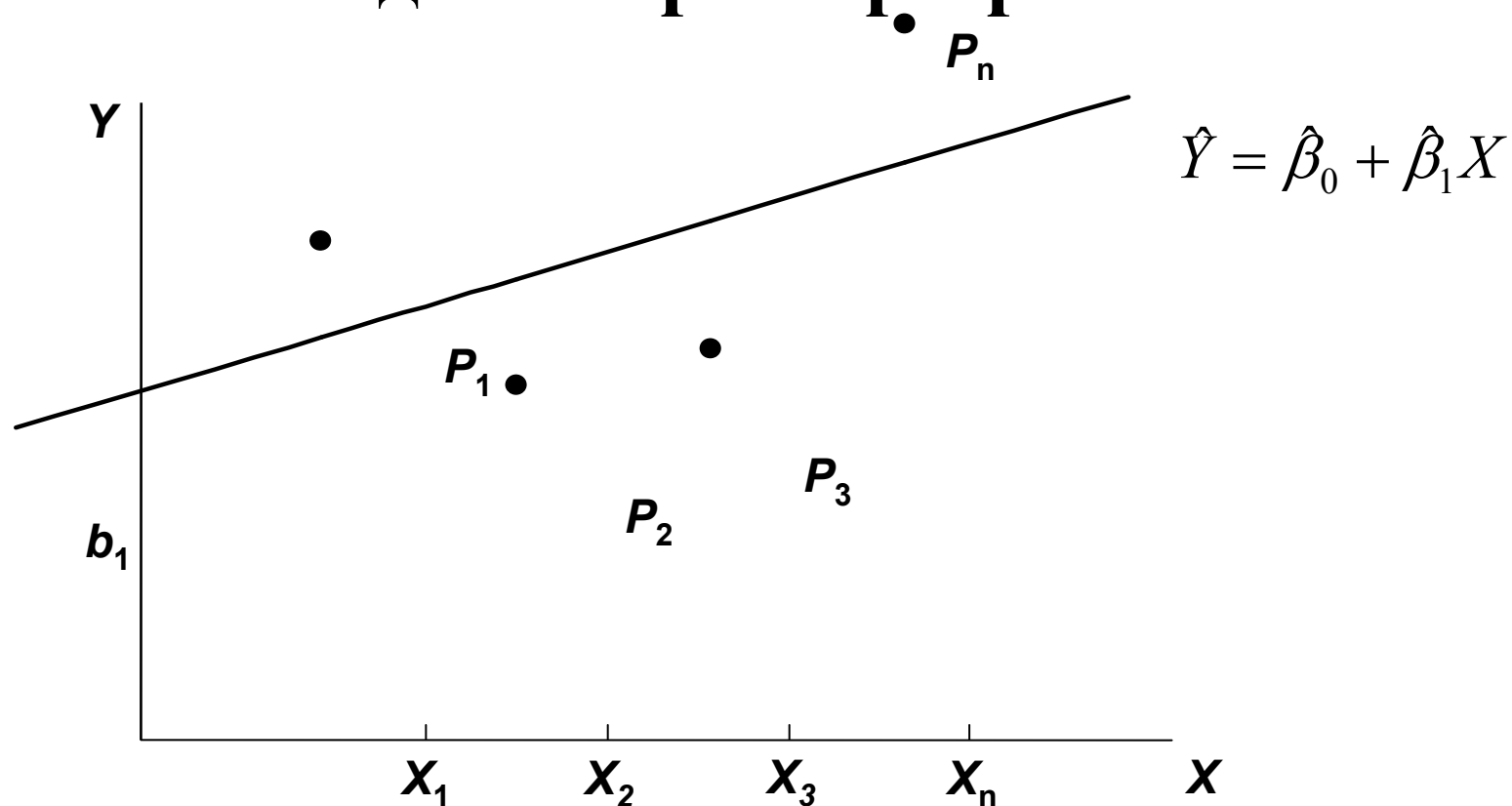
Таким образом, каждое значение переменной Y можно разделить на две части: детерминированную, $\beta_0 + \beta_1 X$, и случайную ε .

Модель парной регрессии



Но на практике мы не имеем линии $Y = \beta_0 + \beta_1 X$, а имеем только n пар наблюдений.

Модель парной регрессии



По n парам наблюдений мы должны построить оценки параметров β_0 и β_1 (соответственно $\hat{\beta}_0, \hat{\beta}_1$).

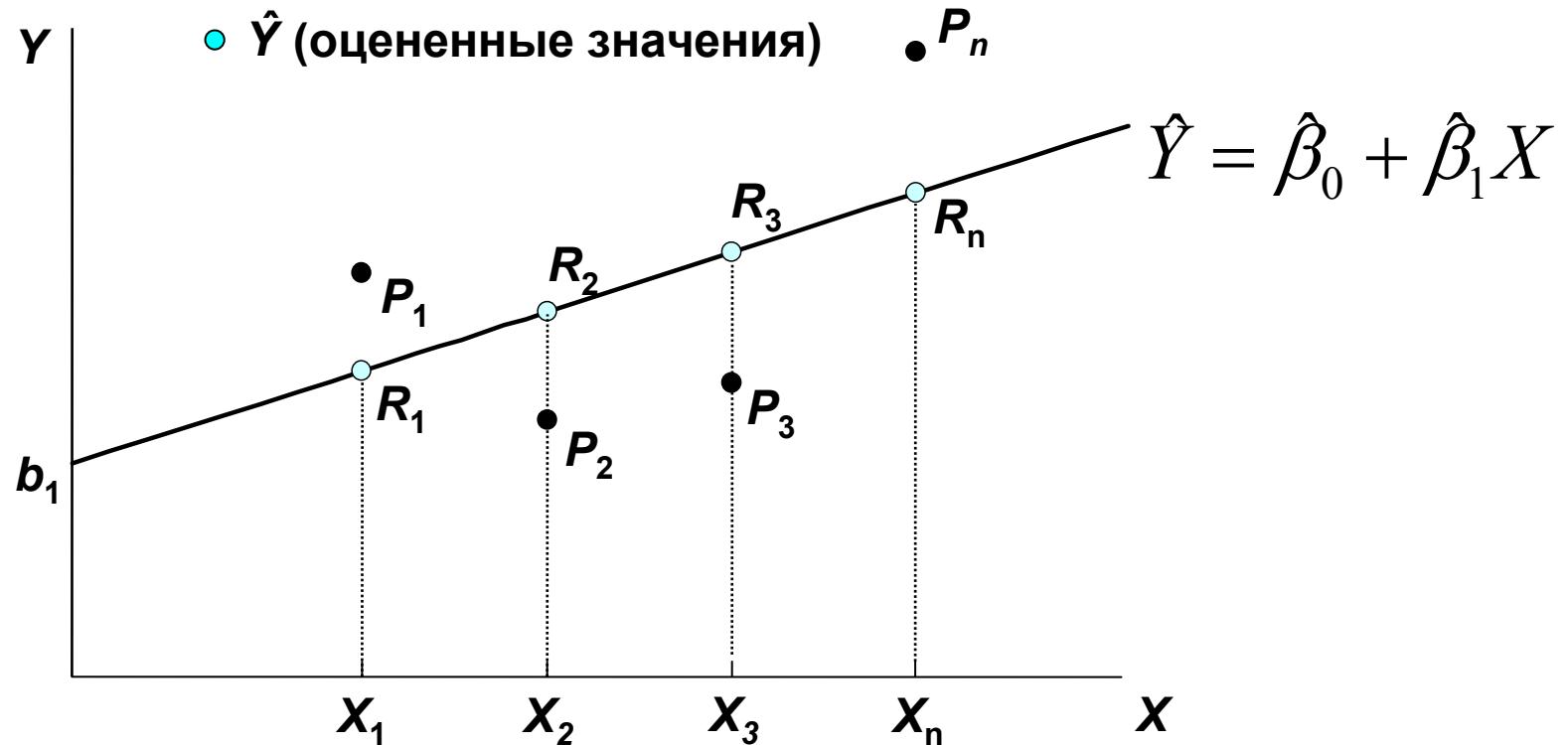
Тогда линия $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

будет являться аппроксимацией линии $Y = \beta_0 + \beta_1 X$.

Модель парной регрессии

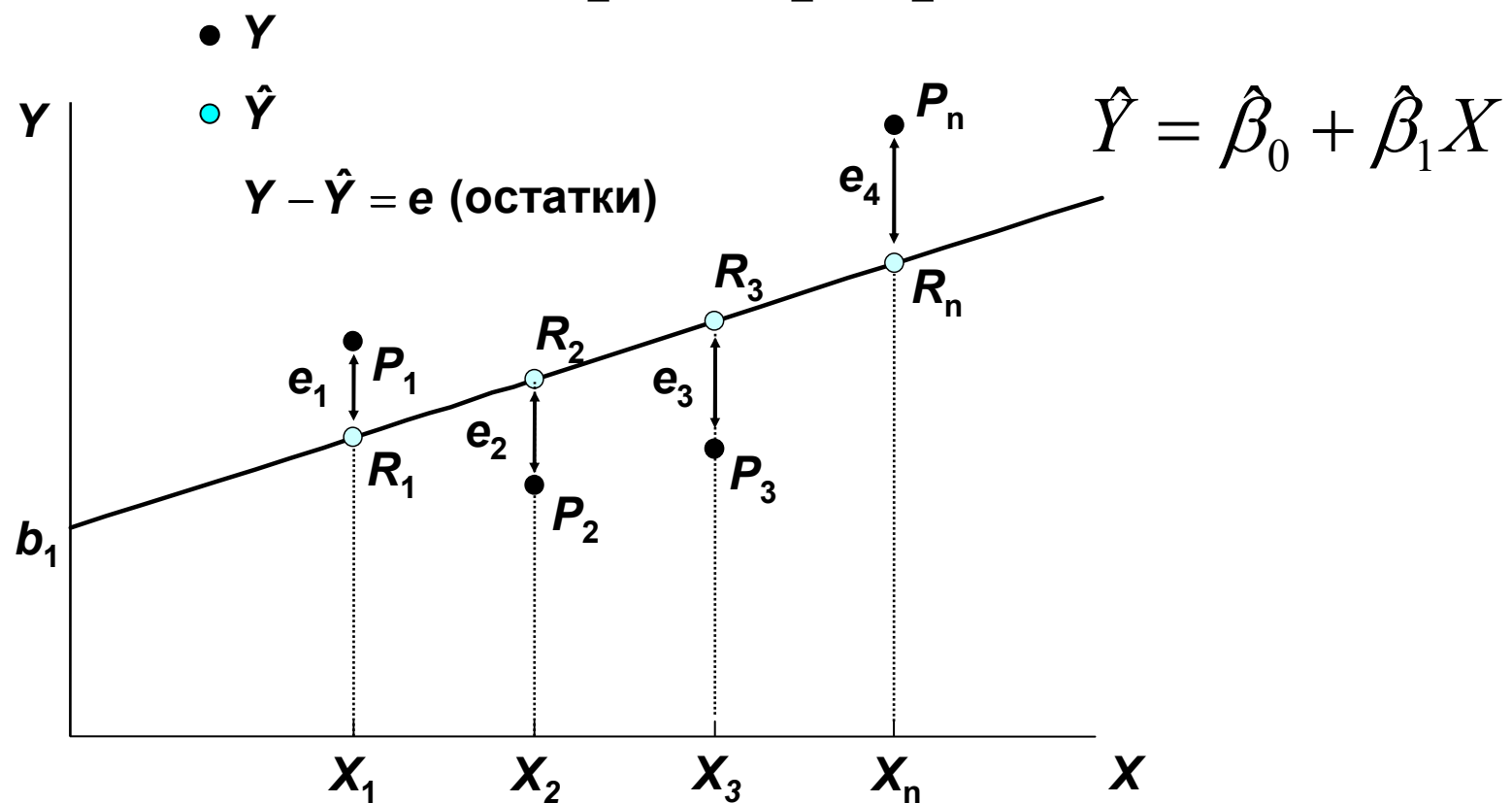
• Y (реальные значения)

• \hat{Y} (оцененные значения) • P_n



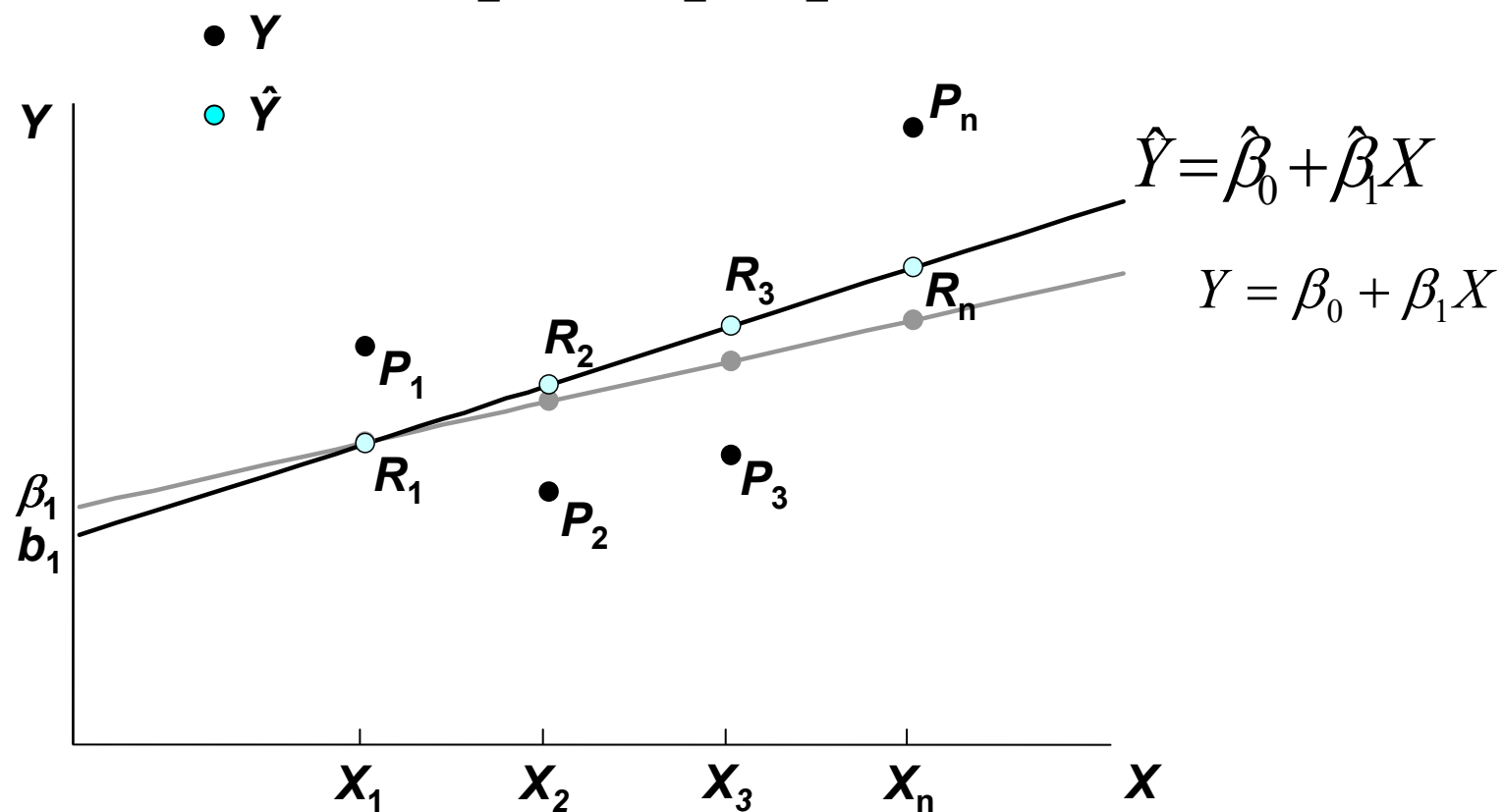
На рисунке проведена линия выборочной регрессии, лежащие на ней точки R_i называются оцененными значениями переменной Y .

Модель парной регрессии



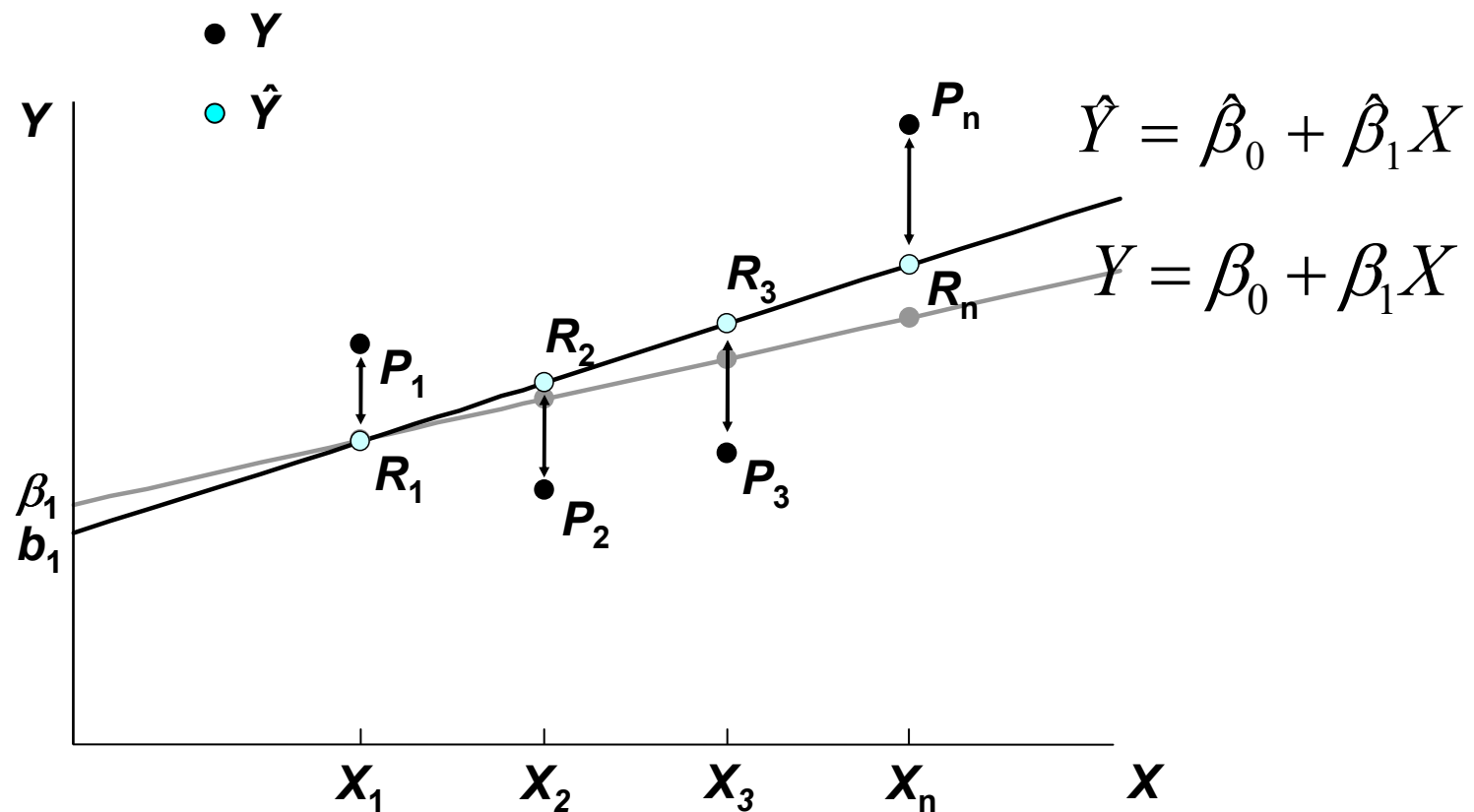
Разности между действительными и оцененными значениями переменной Y называются остатками регрессии.

Модель парной регрессии



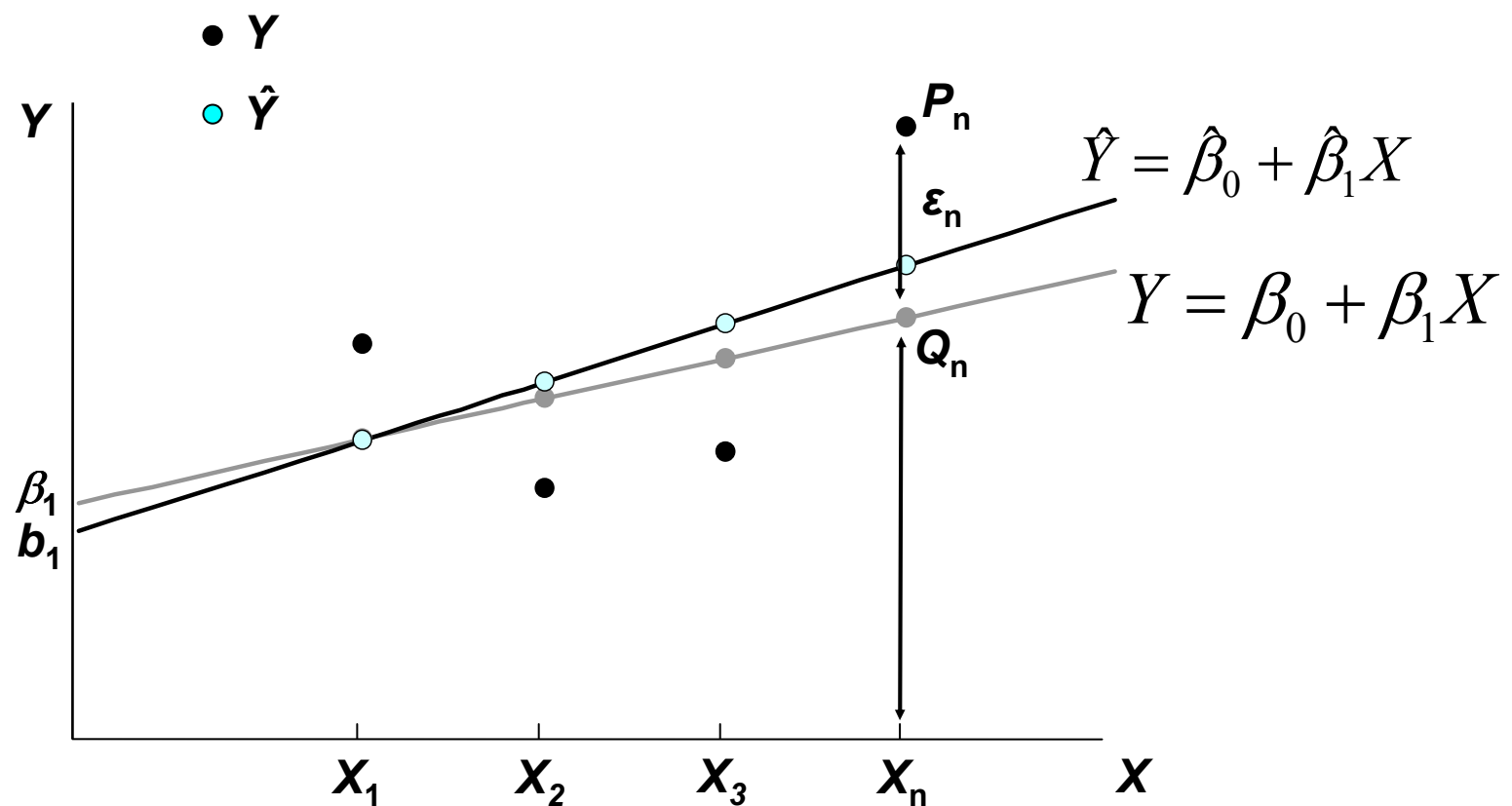
Серым цветом проведена линия теоретической регрессии, а черным – выборочной регрессии.

Модель парной регрессии



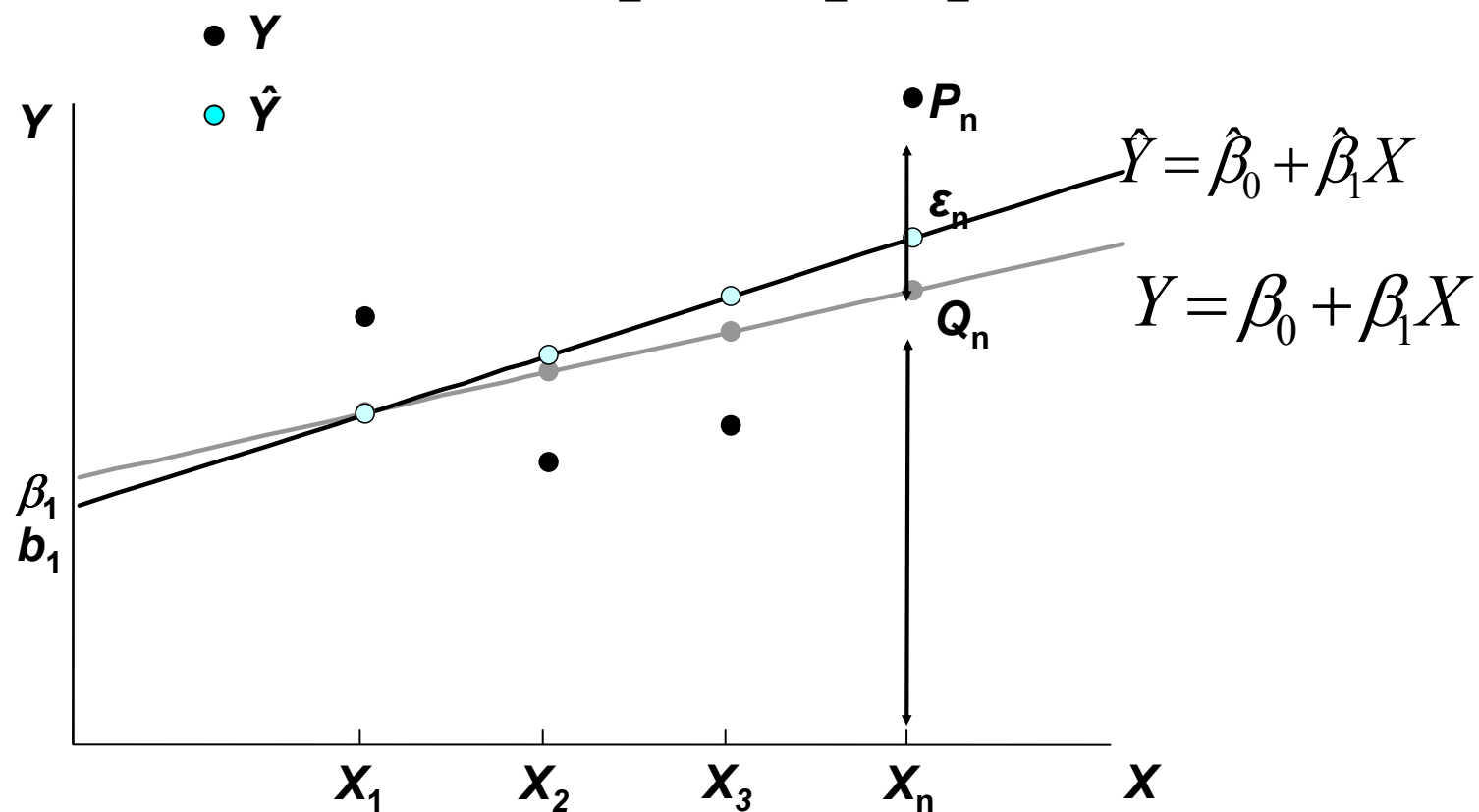
На рисунке изображены остатки e_i (отклонения от линии выборочной регрессии).

Модель парной регрессии



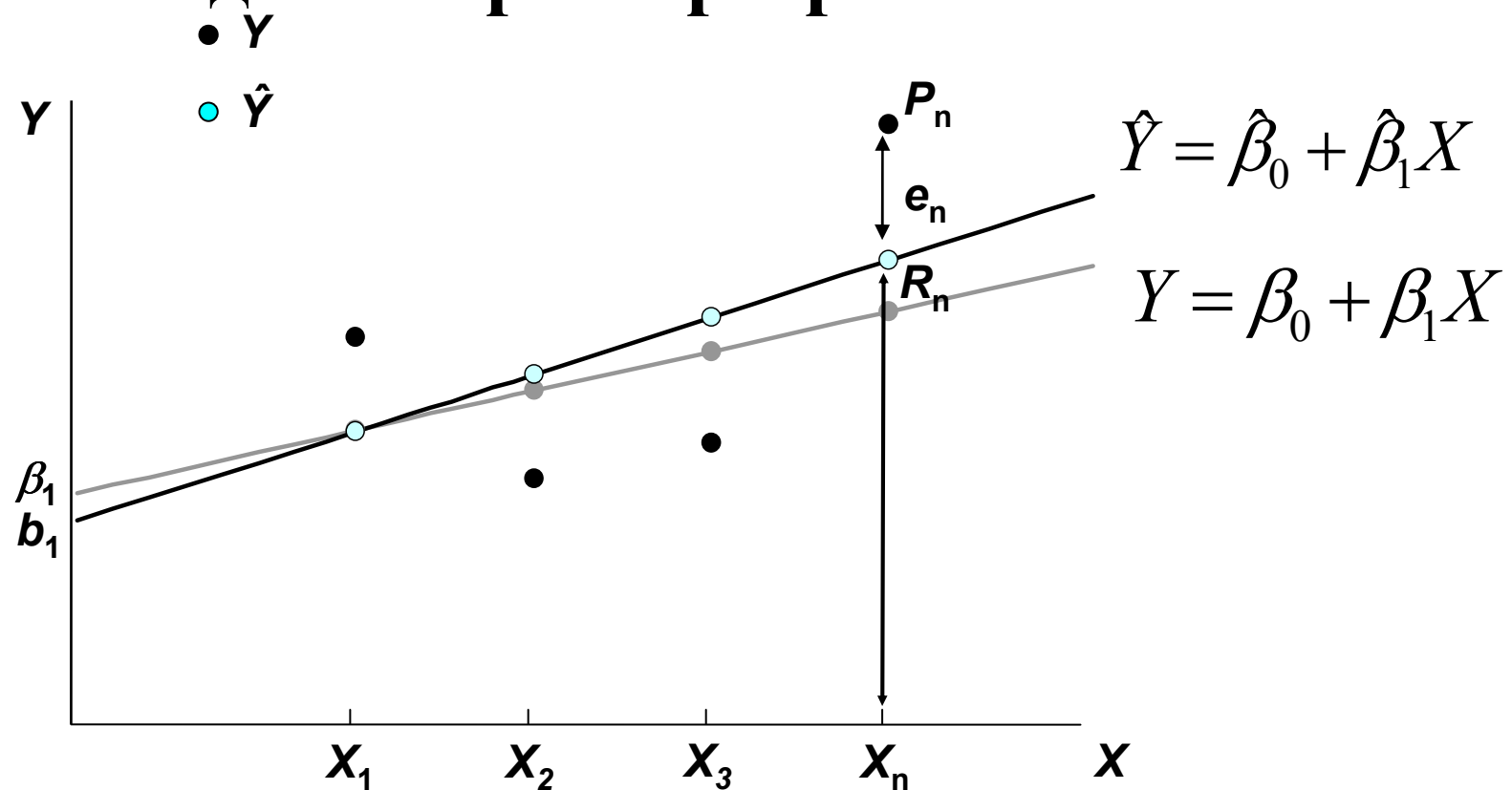
При использовании теоретической регрессии Y разлагается на детерминированную $(\beta_0 + \beta_1 X)$ и случайную(ϵ) части.

Модель парной регрессии



Это разложение является чисто теоретическим (т.к. параметров β_0 и β_1 мы не знаем) и будет использовано при анализе свойств оценок коэффициентов регрессии.

Модель парной регрессии



Другая декомпозиция легко выполнима на практике при известных $\hat{\beta}_0, \hat{\beta}_1$

Оценка коэффициентов выборочной регрессии

Метод наименьших квадратов (МНК) нахождения оценок коэффициентов регрессии состоит в минимизации суммы квадратов остатков регрессии RSS (residual sum of squares).

$$RSS = \sum_{i=1}^n e_i^2 = e_1^2 + \dots + e_n^2$$

МНК

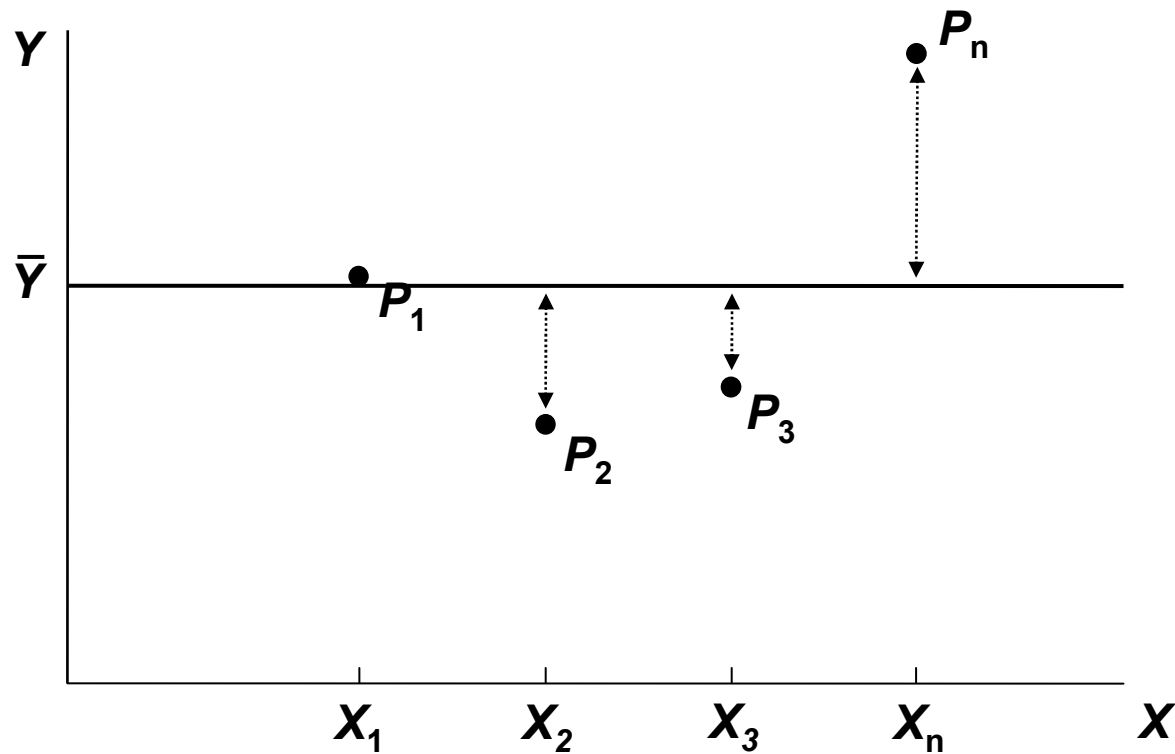
Минимизация RSS (residual sum of squares),

$$RSS = \sum_{i=1}^n e_i^2 = e_1^2 + \dots + e_n^2$$

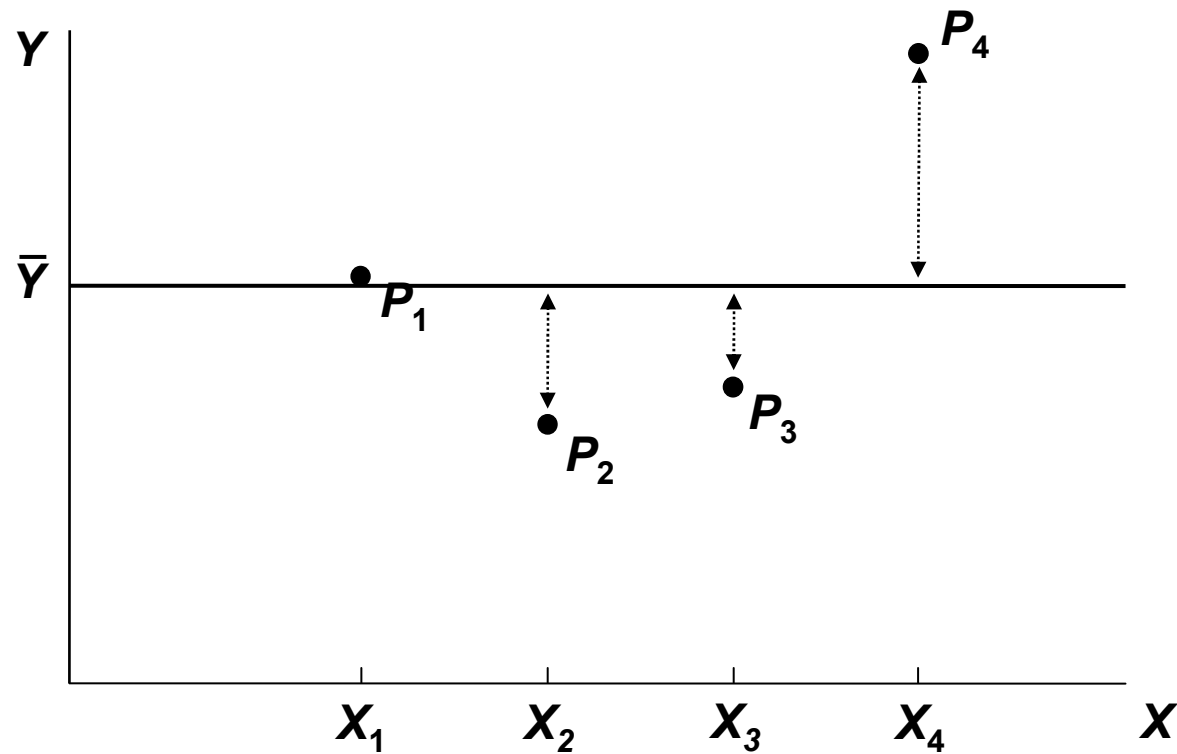
А не суммы остатков

$$\sum_{i=1}^n e_i = e_1 + \dots + e_n$$

Почему минимизируется сумма квадратов остатков, а не сумма остатков?

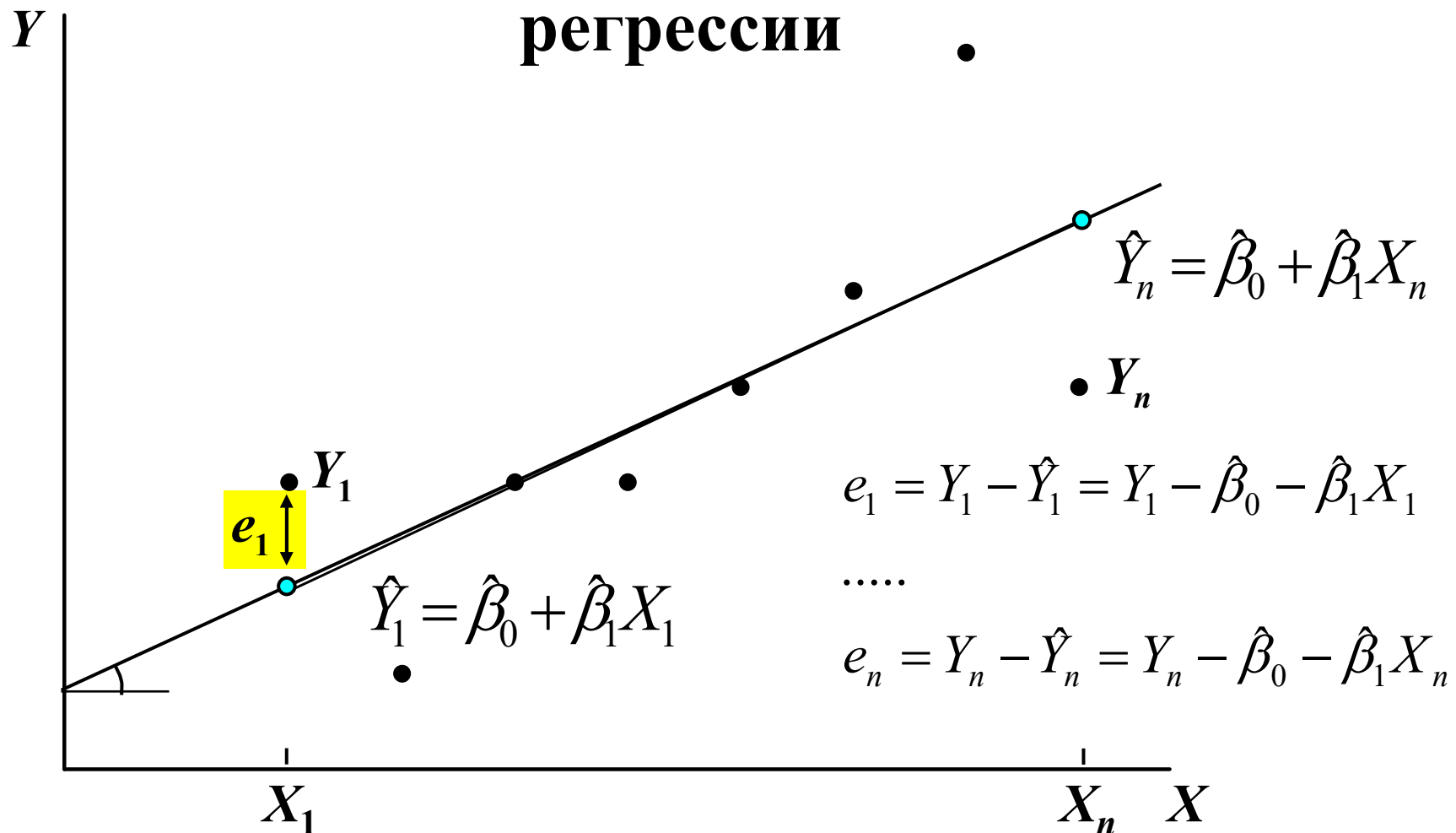


На примере горизонтальной линии легко увидеть, что сумма остатков равна 0, остатки разных знаков компенсируют друг друга, хотя и могут быть велики по абсолютной величине. Это будет иметь место и в общем случае.



МНК является не единственным возможным критерием, но очень удобен для практического применения (обладая и другими замечательными свойствами).

Нахождение оценок коэффициентов парной регрессии



Оцениваемая модель : $Y = \beta_0 + \beta_1 X + u$

Оцененная модель : $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

Выразим остатки регрессии через наблюдения.

Выражение для RSS

$$\begin{aligned}RSS &= e_1^2 + \dots + e_n^2 = (Y_1 - \hat{\beta}_0 - \hat{\beta}_1 X_1)^2 + \dots + (Y_n - \hat{\beta}_0 - \hat{\beta}_1 X_n)^2 \\&= Y_1^2 + \hat{\beta}_0^2 + \hat{\beta}_1^2 X_1^2 - 2\hat{\beta}_0 Y_1 - 2\hat{\beta}_1 X_1 Y_1 + 2\hat{\beta}_0 \hat{\beta}_1 X_1 \\&\quad + \dots \\&\quad + Y_n^2 + \hat{\beta}_0^2 + \hat{\beta}_1^2 X_n^2 - 2\hat{\beta}_0 Y_n - 2\hat{\beta}_1 X_n Y_n + 2\hat{\beta}_0 \hat{\beta}_1 X_n \\&= \sum Y_i^2 + n\hat{\beta}_0^2 + \hat{\beta}_1^2 \sum X_i^2 - 2\hat{\beta}_0 \sum Y_i - 2\hat{\beta}_1 \sum X_i Y_i + 2\hat{\beta}_0 \hat{\beta}_1 \sum X_i\end{aligned}$$

Приводим подобные слагаемые.

Необходимое условие экстремума

$$RSS = \sum Y_i^2 + n\hat{\beta}_0^2 + \hat{\beta}_1^2 \sum X_i^2 - 2\hat{\beta}_0 \sum Y_i - 2\hat{\beta}_1 \sum X_i Y_i + 2\hat{\beta}_0 \hat{\beta}_1 \sum X_i$$

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = 0 \Rightarrow 2n\hat{\beta}_0 - 2\sum Y_i + 2\hat{\beta}_1 \sum X_i = 0$$

$$n\hat{\beta}_0 = \sum Y_i - \hat{\beta}_1 \sum X_i \quad (1)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Приравниваем к 0 частную производную по первой переменной .

Необходимое условие экстремума

$$RSS = \sum Y_i^2 + n\hat{\beta}_0^2 + \hat{\beta}_1^2 \sum X_i^2 - 2\hat{\beta}_0 \sum Y_i - 2\hat{\beta}_1 \sum X_i Y_i + 2\hat{\beta}_0 \hat{\beta}_1 \sum X_i$$

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = 0 \quad \Rightarrow \quad 2\hat{\beta}_1 \sum X_i^2 - 2 \sum X_i Y_i + 2\hat{\beta}_0 \sum X_i = 0$$

Приравниваем к 0 частную производную по второй переменной .

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = 0 \Rightarrow 2\hat{\beta}_1 \sum X_i^2 - 2\sum X_i Y_i + 2\hat{\beta}_0 \sum X_i = 0$$

$$\hat{\beta}_1 \sum X_i^2 - \sum X_i Y_i + \hat{\beta}_0 \sum X_i = 0 \quad (2)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 \sum X_i^2 - \sum X_i Y_i + (\bar{Y} - \hat{\beta}_1 \bar{X}) \sum X_i = 0$$

Уравнения (1) и (2) образуют систему нормальных уравнений.

Делим на 2 и подставляем выражение для $\hat{\beta}_0$ через $\hat{\beta}_1$.

Выражение для $\hat{\beta}_1$

$$\hat{\beta}_1 \sum X_i^2 - \sum X_i Y_i + (\bar{Y} - \hat{\beta}_1 \bar{X}) \sum X_i = 0$$

$$\hat{\beta}_1 \left(\sum X_i^2 - n\bar{X}^2 \right) = \sum X_i Y_i - n\bar{X}\bar{Y}$$

$$\hat{\beta}_1 = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum (X_i - \bar{X})^2}$$

$$\hat{\beta}_1 = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$

Альтернативное выражение для $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum X_i Y_i - n \bar{X} \bar{Y}}{\sum (X_i - \bar{X})^2}$$

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2},$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}, x_i = X_i - \bar{X}, \quad y_i = Y_i - \bar{Y}$$

Разделим числитель и знаменатель на n-1

$$\hat{\beta}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

Интерпретация оценок коэффициентов

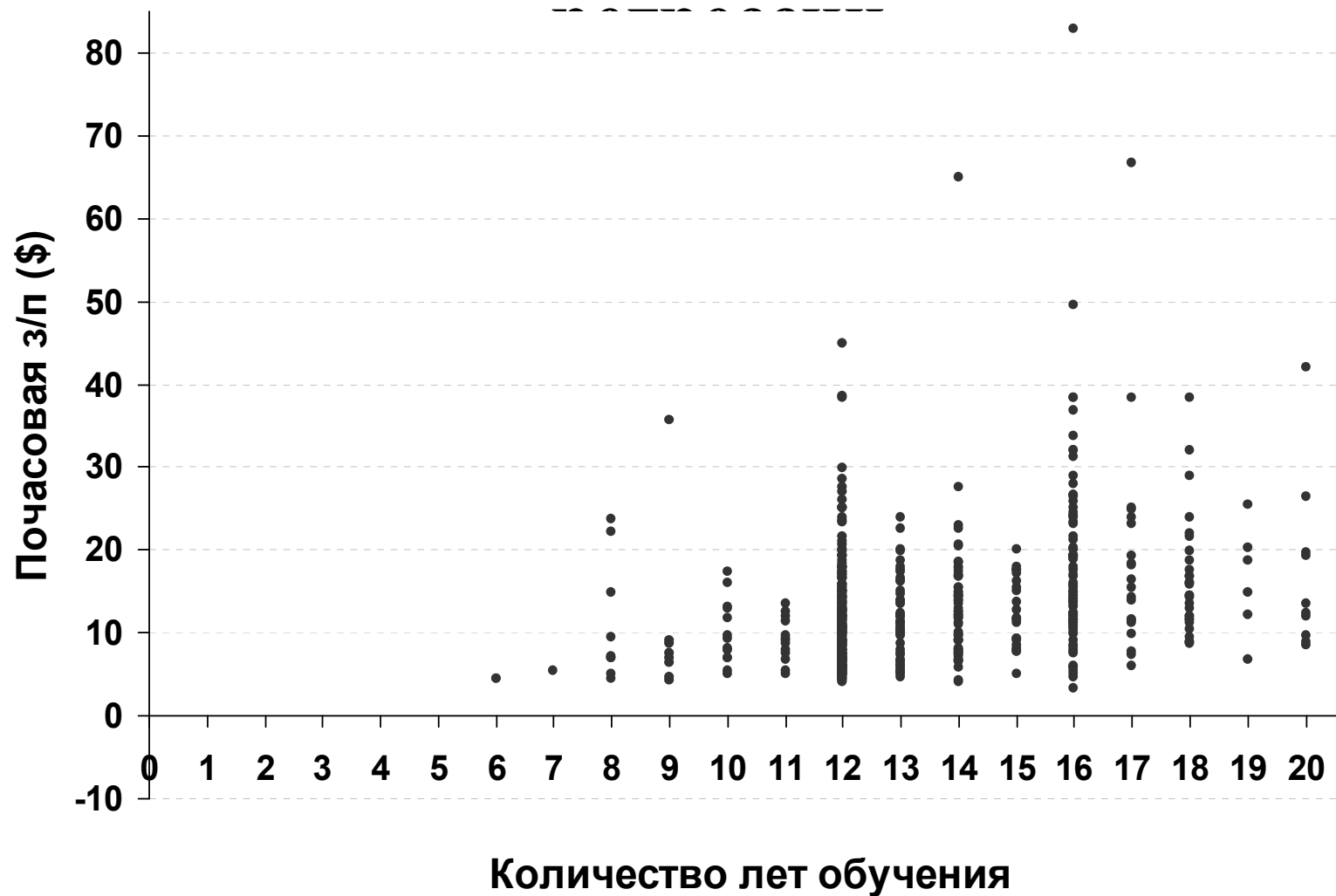
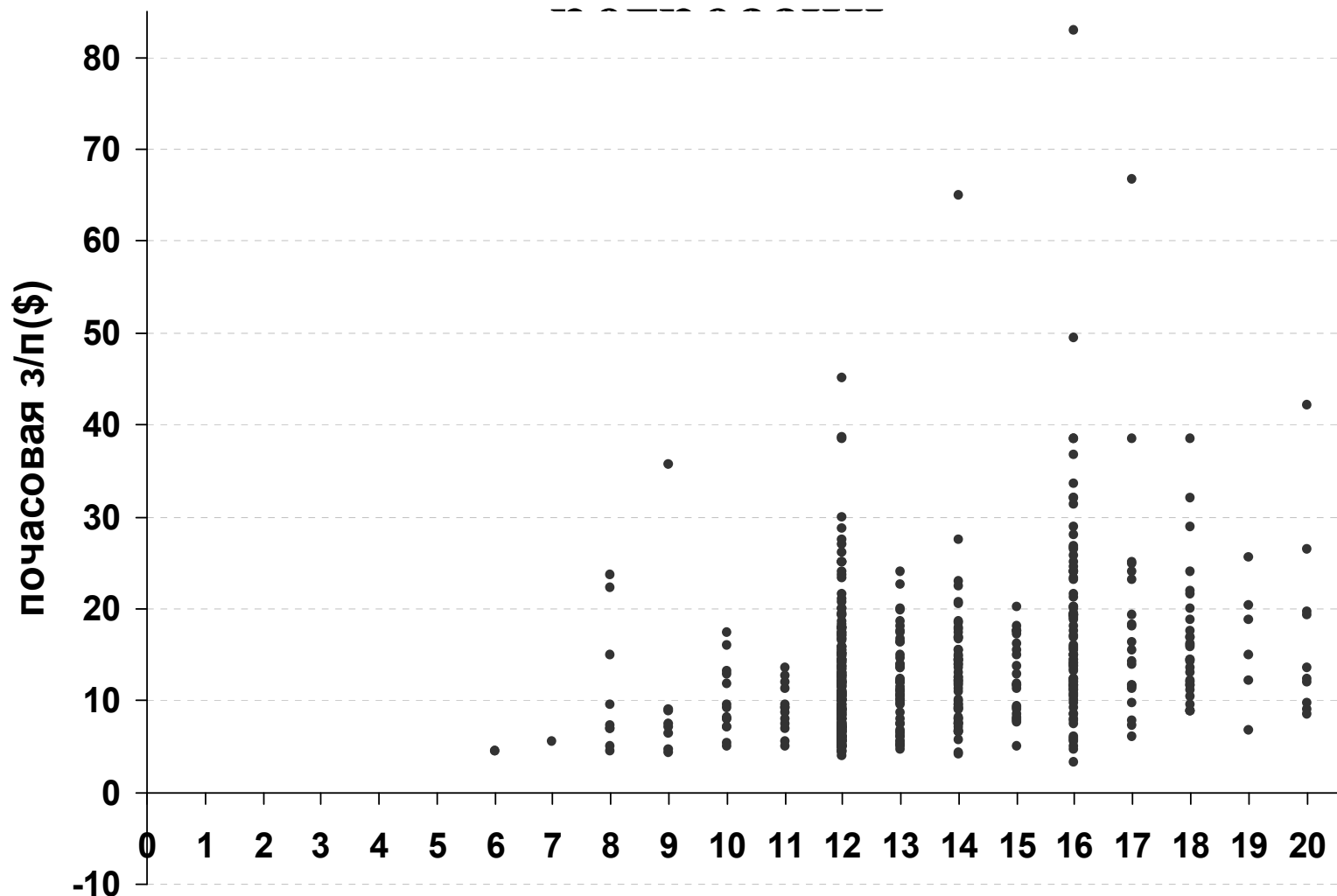


Диаграмма рассеяния отражает зависимость почасовой з/п в 1994 г. от длительности обучения для 570 индивидов из National Longitudinal Survey of Youth.

Интерпретация оценок коэффициентов



6 – 12 лет обучения – школьное образование (неполное или полное),

13 – 15 лет обучения – колледж, 16 – 18 лет – магистратура, 18 – 20 лет – докторантура.

Интерпретация оценок коэффициентов регрессии

```
. reg EARNINGS S
```

Source		SS	df	MS	Number of obs = 570		
-----+-----					F(1, 568) = 65.64		
Model		3977.38016	1	3977.38016	Prob > F = 0.0000		
Residual		34419.6569	568	60.5979875	R-squared = 0.1036		
-----+-----					Adj R-squared = 0.1020		
Total		38397.0371	569	67.4816117	Root MSE = 7.7845		
-----+-----							
EARNINGS		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----							
S		1.073055	.1324501	8.102	0.000	.8129028	1.333206
_cons		-1.391004	1.820305	-0.764	0.445	-4.966354	2.184347
-----+-----							

Для оценки регрессии используется статистический пакет Stata.

Интерпретация оценок коэффициентов регрессии

```
. reg EARNINGS S
```

Source	SS	df	MS	Number of obs	=	570
Model	3977.38016	1	3977.38016	F(1, 568)	=	65.64
Residual	34419.6569	568	60.5979875	Prob > F	=	0.0000
				R-squared	=	0.1036
				Adj R-squared	=	0.1020
				Root MSE	=	7.7845
Total	38397.0371	569	67.4816117			

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	1.073055	.1324501	8.102	0.000	.8129028	1.333206
_cons	-1.391004	1.820305	-0.764	0.445	-4.966354	2.184347

В первой колонке – названия переменных, во второй колонке – оценки коэффициентов регрессии.

Интерпретация оценок коэффициентов регрессии

```
. reg EARNINGS S
```

Source	SS	df	MS	Number of obs	=	570
Model	3977.38016	1	3977.38016	F(1, 568)	=	65.64
Residual	34419.6569	568	60.5979875	Prob > F	=	0.0000
				R-squared	=	0.1036
				Adj R-squared	=	0.1020
Total	38397.0371	569	67.4816117	Root MSE	=	7.7845

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	1.073055	.1324501	8.102	0.000	.8129028	1.333206
_cons	-1.391004	1.820305	-0.764	0.445	-4.966354	2.184347

Оценка коэффициента перед переменной S равна 1.073, а оценка свободного члена (перед cons) равна -1.391.

Интерпретация оценок коэффициентов



На рисунке изображена линия выборочной регрессии.

Интерпретация оценок коэффициентов



S измеряется в годах, $EARNINGS$ в долларах в час. Интерпретация оценки коэффициента наклона: каждый дополнительный год обучения увеличивает почасовую з/п на \$1.07.

Интерпретация оценок коэффициентов



Интерпретация константы, состоящая в том, что индивидуум, не имеющий образования, должен доплачивать за возможность работать, не имеет смысла.

Интерпретация оценок коэффициентов



Однако экстраполяция проведена только для проучившихся более 6 лет, свободный член в данном примере не имеет содержательной экономической интерпретации.