Парный регрессионный анализ

Модель парной линейной регрессии

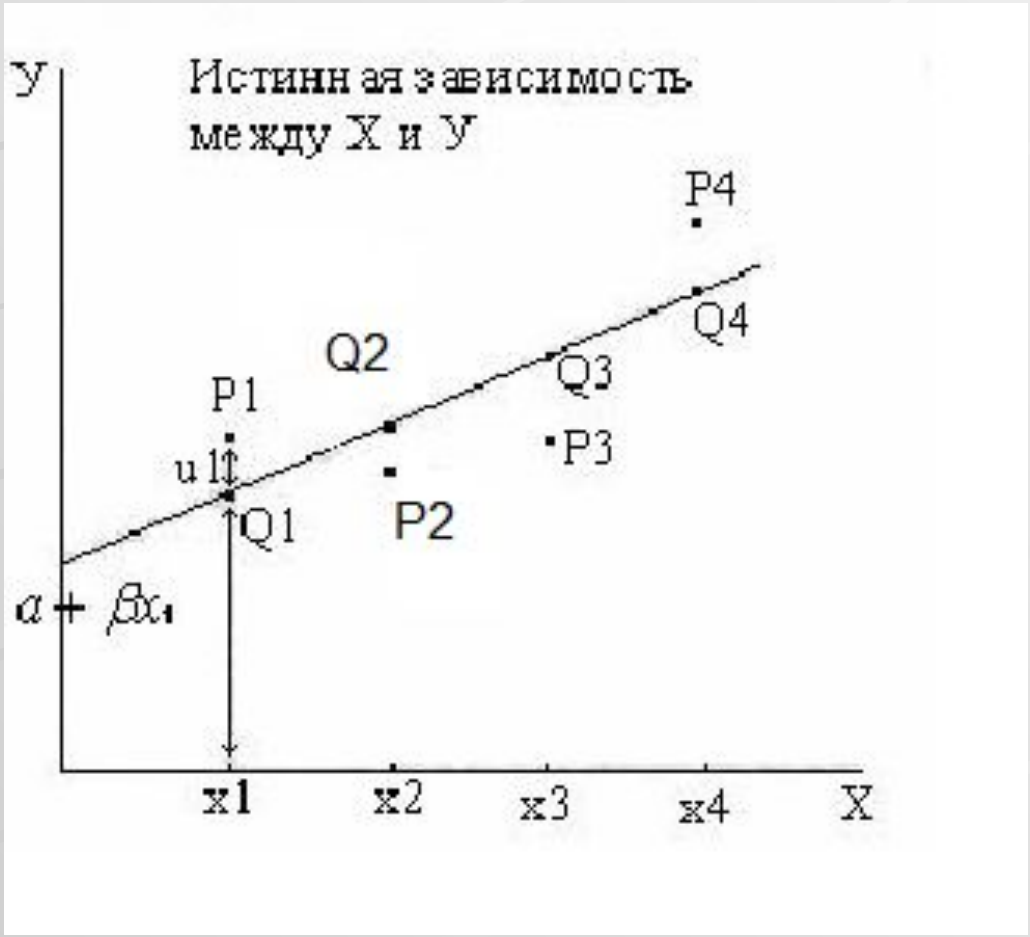
- Коэффициент корреляции показывает, что две переменные связаны друг с другом, однако он не дает представления о том, каким образом они связаны.
- Рассмотрим более подробно те случаи, в которых мы предполагаем, что одна переменная зависит от другой.

- Начнем с простейшей модели - модели **парной линейной регрессии**:

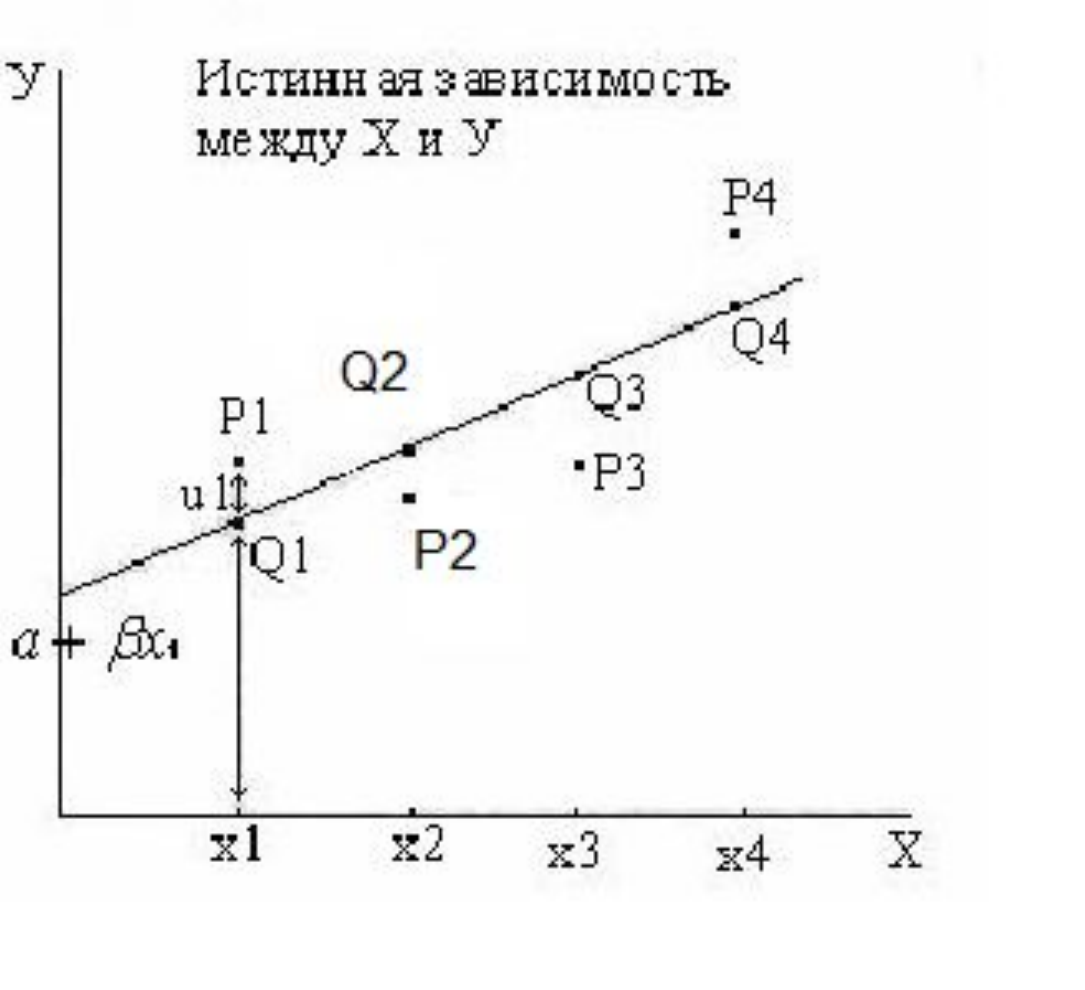
$$y = \alpha + \beta x + u.$$

- Величина y, рассматриваемая как **зависимая** переменная, состоит из двух составляющих:
- **неслучайной составляющей α+βx**, где x выступает как объясняющая (независимая) переменная, а постоянные величины α и β - как параметры уравнения;
- **случайного члена u**.

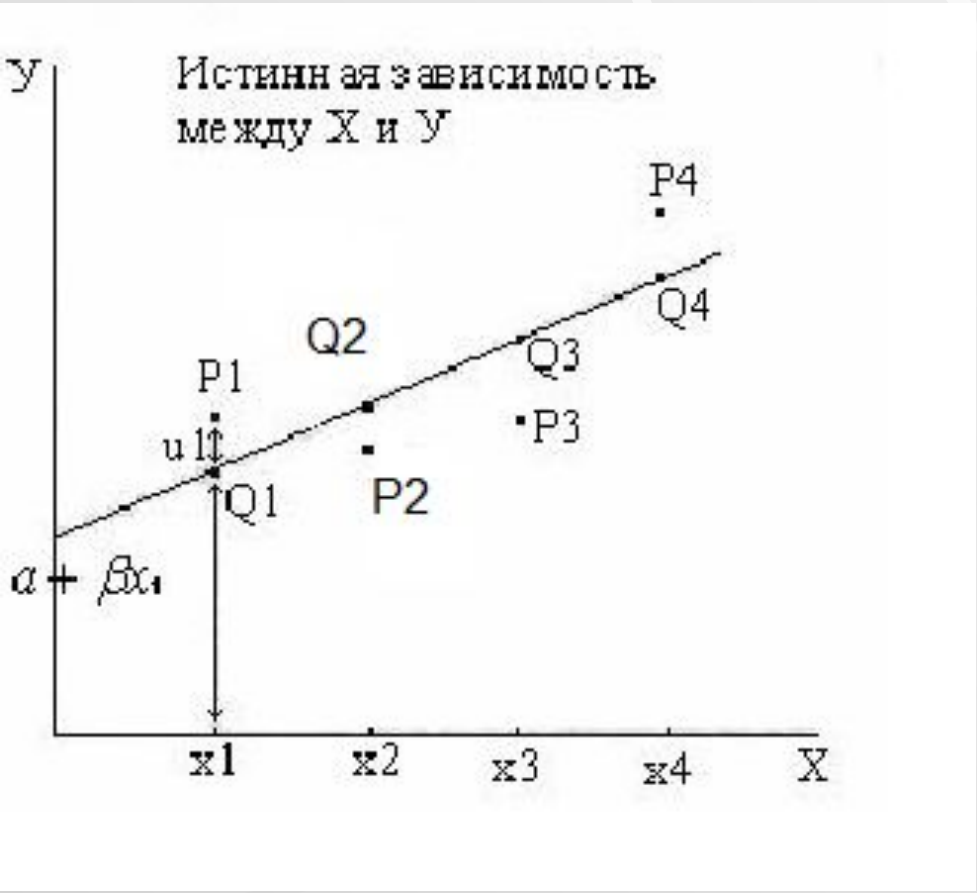
- На рисунке показано, как комбинация этих двух составляющих определяет величину y.



- x1, x2, x3, x4 - это четыре гипотетических значения объясняющей переменной.
- Если бы соотношение между y и x было **линейным**, то значения **y** были бы представлены точками Q1, Q2, Q3, Q4 на прямой.



- Случайный член положителен в первом и четвертом наблюдениях и отрицателен в двух других.
- Если отметить на графике реальные значения y при соответствующих значениях x, то получим точки P1, P2, P3, P4
- Фактические значения α и β и, следовательно, положения точек Q неизвестны.



- Задача регрессионного анализа состоит в получении оценок α и β и, следовательно, в определении положения прямой по точкам P.
- Если бы случайный член отсутствовал, то точки P совпали бы с точками Q и точно показали бы положение прямой.
- В этом случае достаточно было бы просто построить эту прямую и определить значения α и β.

•Почему же существует случайный член?

Причины существования случайного члена

- 1. **Невключение объясняющих переменных**.
- Соотношения между y и x является большим упрощением. В действительности существуют другие факторы, влияющие на y, которые не учтены в формуле $y=\alpha+\beta x+u$.
- 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 психологические факторы.
- Возможно, что существуют также другие факторы, которые можно измерить, но которые оказывают такое слабое влияние, что их не стоит учитывать.

- Кроме того, могут быть факторы, которые являются существенными, но которые мы из-за отсутствия опыта таковыми не считаем.
- Объединив все эти составляющие, мы получаем то, что обозначено как **u**.
- Если бы мы точно знали, какие переменные присутствуют здесь, и имели возможность точно их измерить, то могли бы включить их в уравнение и исключить соответствующий элемент из случайного члена.
- Проблема состоит в том, что мы никогда не можем быть уверены, что входит в данную совокупность, а что - нет.

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

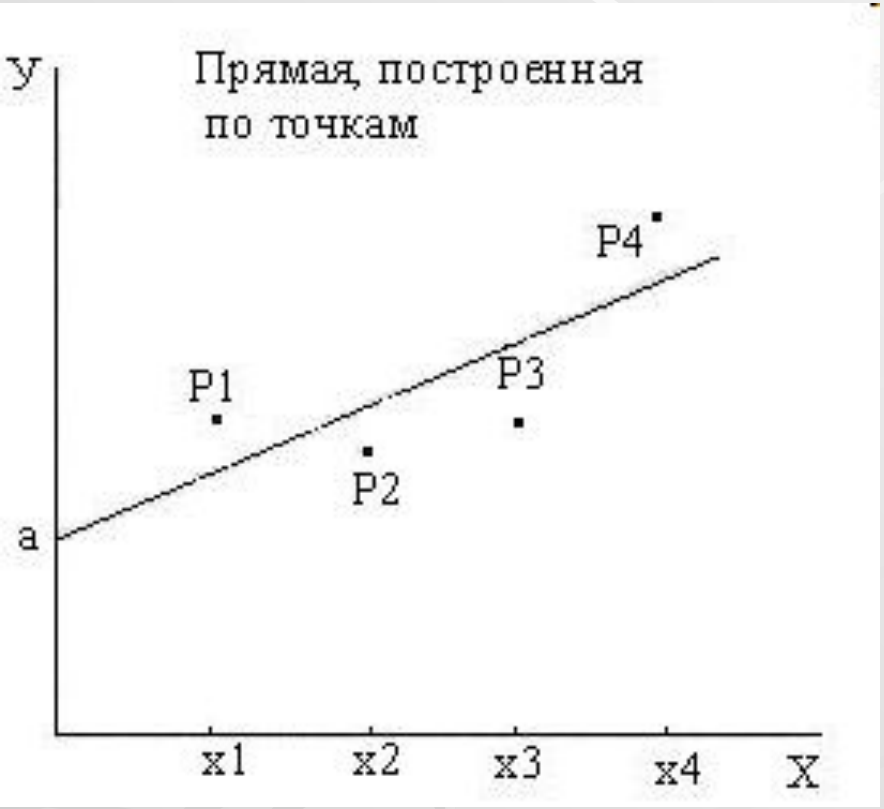
- **3. Неправильное описание структуры модели**.
- Структура модели может быть описана неправильно или не вполне правильно.
- Например, если зависимость относится к данным о временном ряде, то значение y может зависеть не от фактического значения x, а от значения, которое ожидалось в предыдущем периоде.
- 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.
- 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.

- **Ошибки измерения.**
- Если в измерении одной или более взаимосвязанных переменных имеются ошибки, то наблюдаемые значения не будут соответствовать точному соотношению, и существующее расхождение будет вносить вклад в остаточный член.
- **Остаточный член является суммарным проявлением всех этих факторов.**

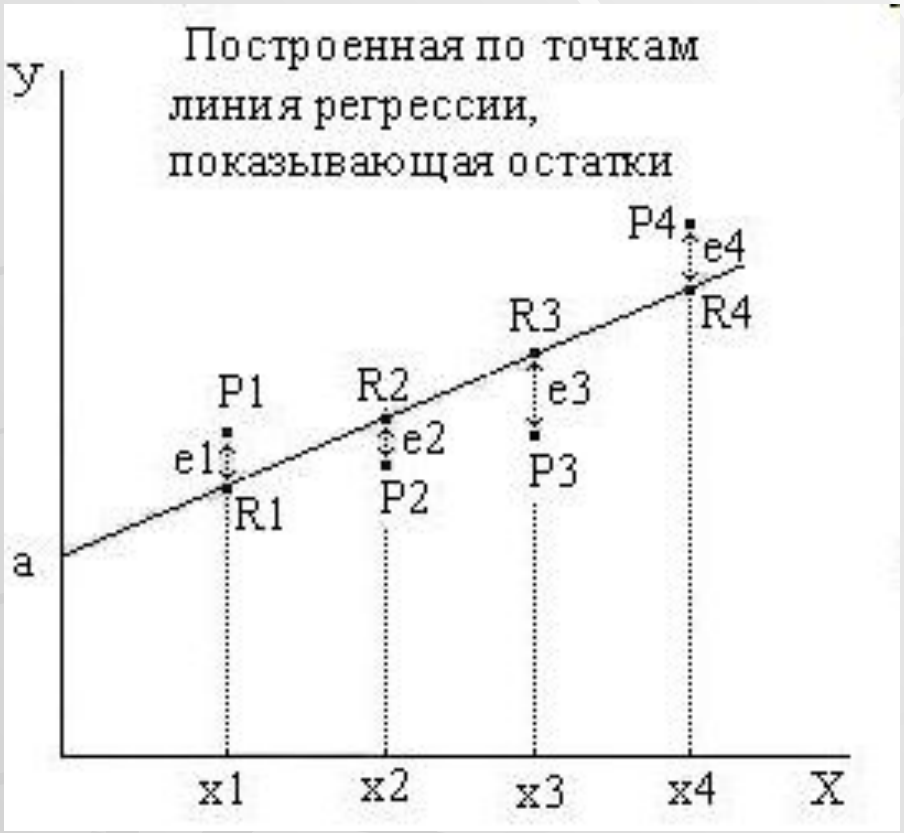
Регрессия по методу наименьший квадратов.

- Допустим, имеется четыре наблюдения для x и y, и поставлена задача - определить значения α и β в уравнении **y=α+βx+u**.
- Можно отложить четыре точки **P** и построить прямую, в наибольшей степени соответствующую этим точкам



- Отрезок, отсекаемый прямой на оси у, представляет собой оценку **α** и обозначен **a**, а угловой коэффициент прямой представляет собой оценку **β** и обозначен **b**.
- Невозможно рассчитать истинные значения **α** и **β,** можно получить только оценки, и они могут быть хорошими или плохими

- **Способ достаточно точной оценки алгебраическим путем – Метод наименьших квадратов**

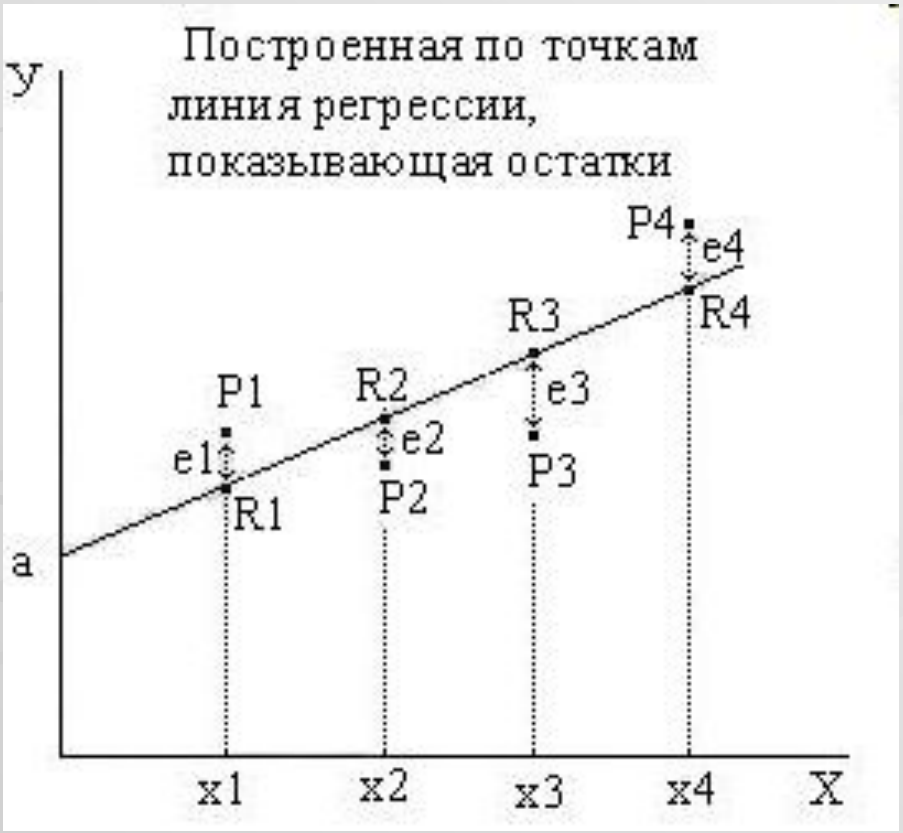


- Первым шагом является определение остатка для каждого наблюдения.
- На рисунке при x=x1 соответствующей ему точкой на линии регрессии будет R1 со значением , $\hat{y}_1$

вместо фактически наблюдаемого значения **y1**.

Величина $\hat{y}_1$ – *расчетное значение*

Разность между фактическим и расчетным значениями $y_1 - \hat{y}_1$ *это остаток* e_1



- Соответственно, для других наблюдений остатки будут обозначены как e_1, e_2, e_3 и e_4.

- Очевидно, что мы хотим построить линию регрессии таким образом, чтобы остатки были минимальными.
- Очевидно также, что линия, строго соответствующая одним наблюдениям, не будет соответствовать другим, и наоборот.
- Необходимо выбрать такой критерий подбора, который будет одновременно учитывать величину всех остатков.

- Одним из способов решения поставленной проблемы состоит в минимизации суммы квадратов остатков S — **метод наименьших квадратов МНК**.
- Для рисунка верно такое соотношение: $S=e_1^2+e_2^2+e_3^2+e_4^2$.

- Величина **S** будет зависеть от выбора **a** и **b**, так как они определяют положение линии регрессии.
- Чем меньше **S**, тем строже соответствие. Если S=0, то получено абсолютно точное соответствие.
- В этом случае линия регрессии будет проходить через все точки, однако, вообще говоря, это невозможно из-за наличия случайного члена.

- Рассмотрим случай, когда имеется **n** наблюдений двух переменных **x** и **y**.
- Предположив, что **y** зависит от **x**, мы хотим подобрать уравнение

$$\hat{y} = a + bx$$

Вывод выражений для a и b

- Выразим квадрат i-го остатка через **a** и **b** и наблюдения значений **x** и **y**:

$$e_i^2 = (y_i - \hat{y}_i)^2 = (y_i - a - bx_i)^2 = y_i^2 + a^2 + a^2 x_i^2 - 2ay_i + 2abx_i - 2x_i y_i$$

Суммируя по всем n наблюдениям, запишем S в виде:

$$S = \sum (y_i)^2 + na^2 + b^2 \sum (x_i)^2 - 2a \sum y_i + 2ab \sum x_i - 2b \sum x_i y_i$$

- Данное выражение для **S** является квадратичной формой по **a** и **b**, и ее коэффициенты определяются выборочными значениями **x** и **y**

- Мы можем влиять на величину **S**, только задавая **a** и **b**.
- Значения **x** и **y**, которые определяют положение точек на диаграмме рассеяния, уже не могут быть изменены после того, как мы взяли определенную выборку.
- Условия первого порядка для минимума, принимают вид:

$$\frac{d}{da}S = 2an - 2\sum y_i + 2b\sum x_i = 0$$

$$\frac{d}{db}S = 2b\sum (x_i)^2 + 2a\sum x_i - 2\sum x_i y_i = 0$$

Эти уравнения называют нормальными уравнениями для коэффициентов регрессии.

- Первое уравнение позволяет выразить **a**

$$2an - 2n\bar{y} + 2bn\bar{x} = 0$$

Следовательно $$a = \bar{y} - b\bar{x}$$

- Подставив выраженное **a** во второе уравнение, затем поделив на **2n** и перегруппировав, получим: **bVar(x)=Cov(x,y),** и таким образом получим уравнение:

b=Cov(x,y)/Var(x).

- Существуют и другие достаточно разумные решения, однако при выполнении определенных условий метод наименьших квадратов дает несмещенные и эффективные оценки **α** и **β**.

Качество оценки: коэффициент детерминации R^2

- Цель регрессивного анализа состоит в объяснении поведения зависимой переменной **y**.
- В любой данной выборке **y** оказывается сравнительно низким в одних наблюдениях и сравнительно высоким - в других.
- Разброс значений **y** в любой выборке можно суммарно описать с помощью выборочной дисперсии **Var(y).**

- В парном регрессионном анализе мы пытаемся объяснить поведение **y** путем определения регрессионной зависимости **y** от соответственно выбранной независимой переменной **x**.
- После построения уравнения регрессии мы можем разбить значение $\mathbf{y_i}$ в каждом наблюдении на две составляющих -

$$y_i = \hat{y}_i + e_i$$

Величина $\hat{y}_i$ – расчетное значение y в наблюдении i

- Это то значение, которое имел бы **y** при условии, что уравнение регрессии было бы правильным, и отсутствии случайного фактора.

Величина **y** спрогнозированная по значению **x** в данном наблюдении.

Тогда остаток $\mathbf{e_i}$ есть расхождение между фактическим и спрогнозированным значениями величины **y**.

Это та часть **y**, которую мы не можем объяснить с помощью уравнения регрессии.

$$\mathrm{var}(y) = \mathrm{var}(\hat{y}) + \mathrm{var}(e) + 2\,\mathrm{cov}(\hat{y}, e)$$

$$2\,\mathrm{cov}(\hat{y}, e) = 0 \;\Rightarrow\; \mathrm{var}(y) = \mathrm{var}(\hat{y}) + \mathrm{var}(e)$$

Значит мы можем разложить Var(y) на две части

$\mathrm{var}(\hat{y})$ – часть, которая "объясняется" уравнением регрессии,

и Var(e) - "необъяснимую" часть.

$$R^2 = \frac{\text{var}(\hat{y})}{\text{var}(y)} = 1 - \frac{\text{var}(e)}{\text{var}(y)}$$

Это коэффициент детерминации

Максимальное значение коэффициента R^2 равно единице. Это происходит в том случае, когда линия регрессии точно соответствует всем наблюдениям

$$\textit{Тогда } \text{var}(y) = \text{var}(\hat{y})$$

$$\text{Var(e)} = 0 \text{ и } \text{R}^2 = 1.$$

- Если в выборке отсутствует видимая связь между **y** и **x**, то коэффициент $\mathbf{R^2}$ будет близок к нулю.
- При прочих равных условиях желательно, чтобы коэффициент $\mathbf{R^2}$ был как можно больше.
- В частности, мы заинтересованы в таком выборе коэффициентов **a** и **b**, чтобы максимизировать $\mathbf{R^2}$

- Принцип минимизации суммы квадратов остатков эквивалентен минимизации дисперсии остатков.
- Однако, если мы минимизируем **Var(e),** то при этом в соответствии с $\mathbf{R^2=1-Var(e)/Var(y)}$ автоматически максимизируется коэффициент **R2**.

Пример вычисления коэффициента R2

Уравнение регрессии $\hat{y} = 1{,}6667 + 1{,}5x$

Набл.	x	y	$\hat{y}$	e	$y - \bar{y}$	$\hat{y} - \bar{y}$	$(y - \bar{y})^2$	$(\hat{y} - \bar{y})^2$	e^2
1	1	3	3,1667	-0,1667	-1,6667	-1,5	2,7778	2,25	0,0278
2	3	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
Сумма	**6**	**14**	**14**	**0**			4,6667	4,50	0,1667
Среднее	**2**	**4,6667**	**4,6667**	**0**			1,5556	1,50	0,0556

$R^2 = 1{,}5000/1{,}5556 = 0{,}96$

- Общая сумма квадратов **TSS** – сумма квадратов отклонений **y** от своего среднего значения
- Объясненная сумма квадратов (**ESS**) отклонений – сумма квадратов отклонений **a+bx** от выборочного среднего.
- Необъясненная (остаточная) сумма квадратов отклонений (**RSS**) – сумма квадратов остатков всех наблюдений.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$RSS = \sum_{i=1}^{n} e_i^2$$

Другая формула для коэффициента детерминации

$$R^2 = \frac{ESS}{TSS} \qquad TSS = RSS + ESS$$

Связь между коэффициентом детерминации и коэффициентом корреляции

$$r_{y,\hat{y}} = \frac{\operatorname{cov}(y,\hat{y})}{\sqrt{\operatorname{var}(y)\operatorname{var}(\hat{y})}} = \frac{\operatorname{cov}(\hat{y}+e,\hat{y})}{\sqrt{\operatorname{var}(y)\operatorname{var}(\hat{y})}} = \frac{\operatorname{cov}(\hat{y},\hat{y})+\operatorname{cov}(e,\hat{y})}{\sqrt{\operatorname{var}(y)\operatorname{var}(\hat{y})}} =$$

$$= \frac{\operatorname{var}(\hat{y})}{\sqrt{\operatorname{var}(y)\operatorname{var}(\hat{y})}} = \frac{\sqrt{\operatorname{var}(\hat{y})}}{\sqrt{\operatorname{var}(y)}} = \sqrt{R^2}$$

$$т.к.\ \operatorname{cov}(e,\hat{y}) = 0$$

То есть коэффициент детерминации равен квадрату выборочной корреляции между **y** и **a+bx**

- Чем ближе коэффициент детерминации к 1 , тем ближе выборка к **линии регрессии y=a+bx, а не к истинной прямой**.
- Это один из недостатков МНК.