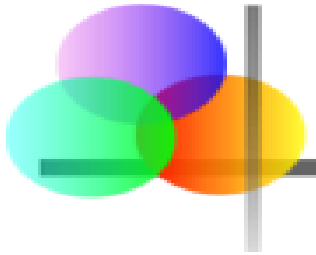


Бизнис статистика

Линеарна регресија

Испитување на врски помеѓу променливи

- Повеќето статистички испитувања се однесуваат на повеќе од едно обележје за дадена популација.

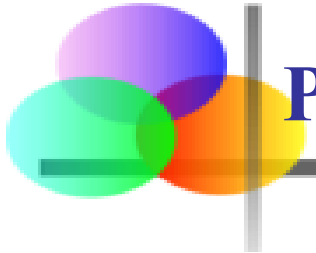
- За да се испита врската помеѓу две или повеќе обележја, потребно е тие обележја да се мерат на иста група единки.

- Природата и јачината на врската помеѓу две или повеќе квантитативни обележја може да се испита со две познати техники во статистиката, *регресиона* и *корелациона анализа*, кои иако се поврзани имаат различни намени.

Регресиона и корелациона анализа

- *Регресионата анализа* се користи за одредување на видот на врската (функционалната поврзаност) на обележјата и главната цел кога се користи овој метод е да се предвиди или процени вредноста на едната променлива за дадена вредност на другата променлива.

- *Корелационата анализа*, од друга страна, е поврзана со мерење на јачината на врската помеѓу променливите, односно го определува степенот на корелација помеѓу променливите.



Регресиона анализа

- Честопати во пракса се сретнуваме со мерења на повеќе обележја на исти единки од популацијата.

- Вакви мерења најчесто се прават за да се провери како промените на одредени обележја (независни променливи) влијаат на промените на обележјето од интерес (зависна променлива).

- Примената на техниките на регресиона анализа е многу широка. Овие техники наоѓаат примена во инженерство, медицински, биолошки и општествени науки, физички, хемиски науки, економија, ...

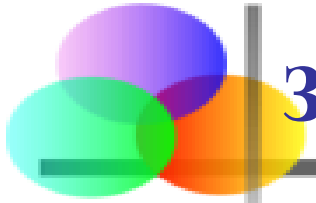
- Регресијата е една од најчесто користените статистички методи.

Регресиона анализа - примена

- Еве неколку примери од примена во кои може да се користи регресијата:
  - Како цената за стан зависи од големината, локацијата, ...
  - Како растењето на растенијата зависи од ѓубривото, квалитетот на почвата, ...
  - Како зависи премијата за осигурување на домот од возраста на сопственикот на куќата, вредноста на домот и неговата содржина, локација, ...
  - Како на потрошувачката на електрична енергија во домаќинството во текот на еден месец влијае просечната надворешна температура.

Терминологија

- Променливата што е предмет на нашата студија се нарекува **зависна променлива**.
- Променливата која можеме да ја контролираме на извесен начин или нејзините вредности да ги бираме на произволен начин велиме дека претставува **независна променлива**.
- Независната променлива ги објаснува или влијае на промените на другата променлива.
- Вообичаено е зависната променлива да се означува со Y, а независната со X. Доколку пак на мерењата на зависната променлива влијаат повеќе независни променливи ги означуваме со X_1, X_2, …, X_n.



Зависна и независни променливи

- Така, ако го разгледаме примерот за зависноста на цената на станот од големината и локацијата, тогаш
  - Зависна променлива е цената на станот,
  - Независни променливи се големината и локацијата.

- Ако се разгледува зависноста на премијата за осигурување на домот од возраста на сопственикот на куќата, вредноста на домот и локацијата, тогаш
  - Зависна променлива е висината на премијата,
  - Независни променливи се возраста на сопственикот, вредноста на домот и локацијата.

Модели на регресија

- Задачата е да се направи математички модел кој ќе ја изрази Y како функција од независните променливи врз база на податоци односно мерења на Y за конкретни вредности на $X_1, X_2, \ldots, X_k$.

- Меѓутоа, треба да се има предвид дека при реални ситуации не можеме да зборуваме за детерминистичка зависност, туку на мерењата често влијаат и случајни фактори кои не можат да се контролираат од истражувачот.

- Статистичките моделите кои се користат за поврзување на зависна променлива Y со независните променливи $X_1, X_2, \ldots, X_k$ се наречени ***модели на регресија.***

Модели на регресија

- Општиот облик на овие модели е

$$Y = f(X_1, X_2, \ldots, X_k) + \varepsilon$$

каде

  - Y е зависна случајна променлива,
  - $X_1, X_2, \ldots, X_n$ се независни променливи
  - ε е компонента на случајна грешка.

- Математичкото очекување на случајната променлива Y за дадени вредности на независните променливи $X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k$ можеме да го запишеме со

$$E(Y \mid x_1, x_2, \ldots, x_k) = f(x_1, x_2, \ldots, x_k).$$

Линеарни модели

- Моделите кои го изразуваат математичкото очекување на случајната променлива Y за дадени вредности на независните променливи $X_1 = x_1$, $X_2 = x_2$, …, $X_k = x_k$ како линеарна функција од множество непознати параметри се познати како *линеарни модели*.

- Моделот на проста линеарна регресија

$$E(Y \mid x) = \beta_0 + \beta_1 x$$

го поврзува Y со вредноста на една независна променлива X и ја дава претпоставката дека математичкото очекување на Y, за дадена вредност $X = x$, графички е права.

Линеарни модели

- Моделите за повеќекратна регресија се слични со простите линеарни регресиони модели освен што тие содржат повеќе членови и може да се искористат за предложување на врски посложени од праволинискиите.

- На пример, претпоставуваме дека средното време потребно за да се изврши процесирање на податоци на даден компјутер се зголемува со зголемувањето на употребата на компјутерот и врската е криволиниска.
  - Ако x_1 е променлива која ја мери употребата на компјутерот, тогаш наместо *праволиниски модел* $E(Y \mid x_1) = \beta_0 + \beta_1 x_1$ можеме да користиме *квадратен модел*

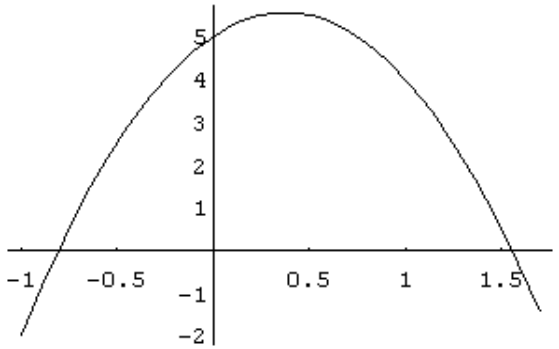
$$E(Y \mid x_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2.$$

Линеарни модели

- Праволинискиот модел $E(Y \mid x) = \beta_0 + \beta_1 x_1$ се нарекува *прво-степен модел*, додека квадратниот модел

$$E(Y \mid x_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2,$$

се нарекува второ-степен модел и графички е парабола.

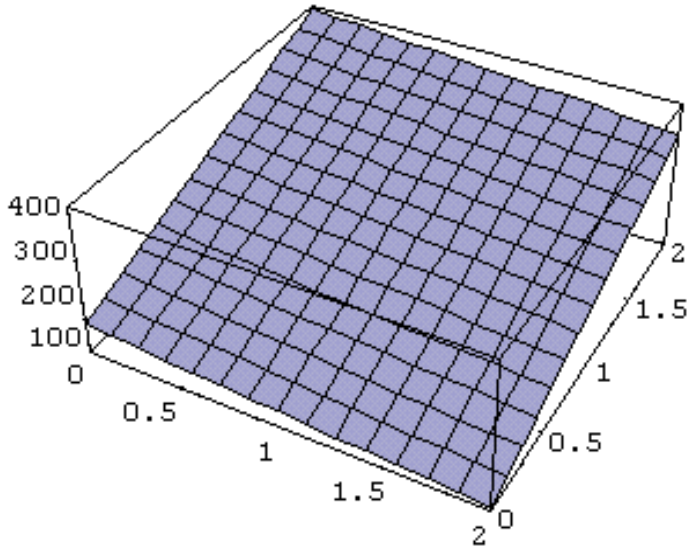


Линеарни модели

- Ако средното време потребно за процесирање на податоци е поврзано и со обемот на работата x_2, тогаш можеме да го вклучиме и x_2 во моделот, па тој добива облик:

$$E(Y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

- Графикот на $E(Y|x_1,x_2)$ како функција од x_1 и x_2 е површина врз (x_1, x_2)-рамнината. На пример, прво-степениот модел претставува рамнина над (x_1, x_2)-рамнината.



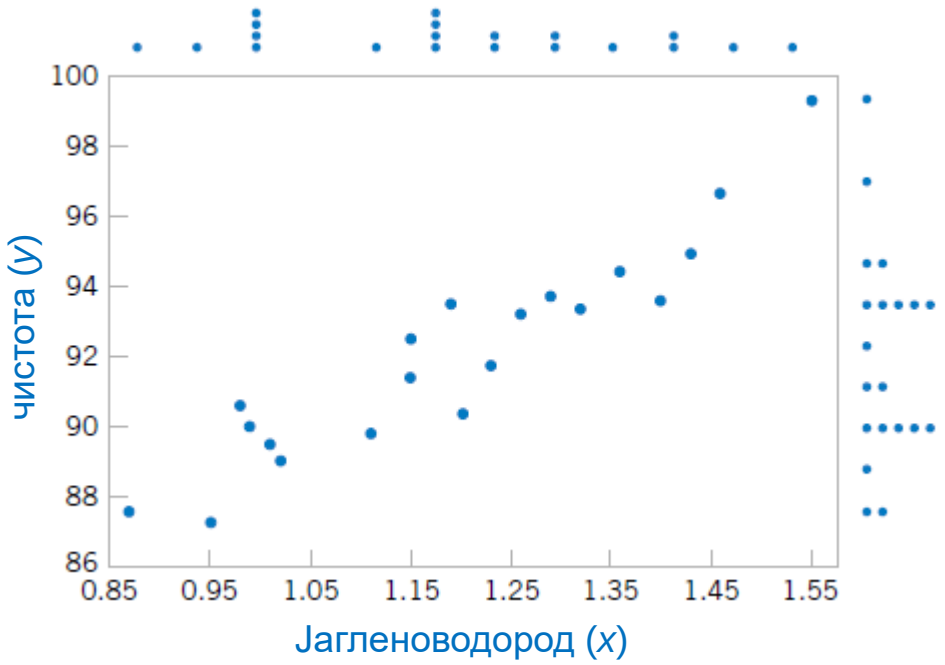
Избор на модел

- Изборот на соодветен регресионен модел за посебна ситуација е многу важно.
- Затоа, прво треба графички да се прикажат резултатите односно точките кои одговараат на измерени вредности на секоја индивидуа.
- Доколку станува збор за само две променливи, вредностите на независната променлива обично се нанесуваат на хоризонталната оска, а вредностите на зависната промелива на вертикалната оска.
- Секоја индивидуа е прикажана како точка во рамнината чии координати се соодветните вредности на независната и зависната променлива.

Избор на модел

- На пример, во табелата се дадени вредностите на x и y, каде y е чистотата на кислород произведен со хемиска дестилација, а x е процентот на јагленоводород што е присутен во главниот кондензатор на единицата за дестилација.
- На сликата е даден соодветниот дијаграм на расејување.

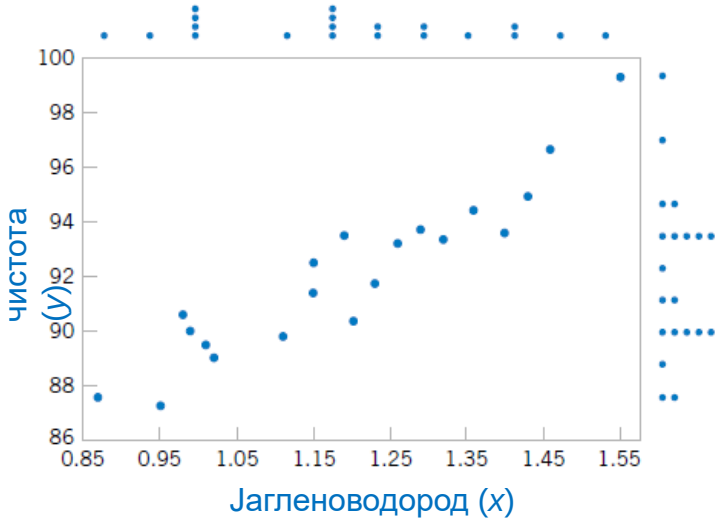
бр. на набљудување	јагленоводород x (%)	чистота y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33



Избор на модел

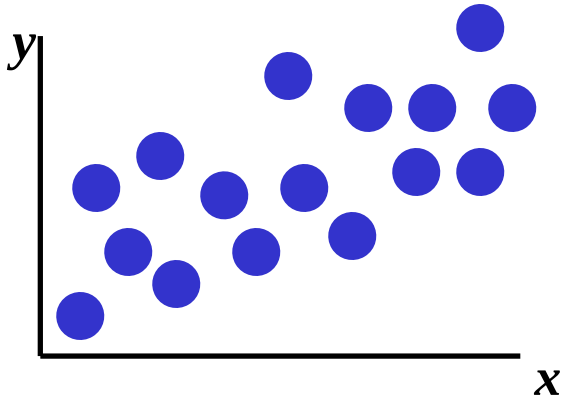
- Проучувањето на овој дијаграм на расејување покажува дека, иако не постои едноставна крива која минува низ сите точки, постои силна индикација дека точките лежат случајно расфрлани околу права.
- Затоа, разумно е да се претпостави дека очекуваната вредност на случајната променлива Y е поврзана со x со линеарната врска:

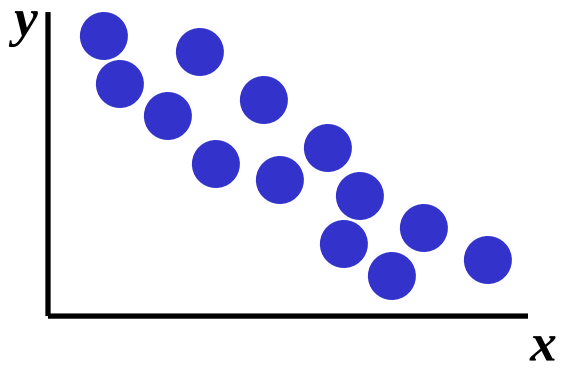
$$E(Y \mid x) = \beta_0 + \beta_1 x$$



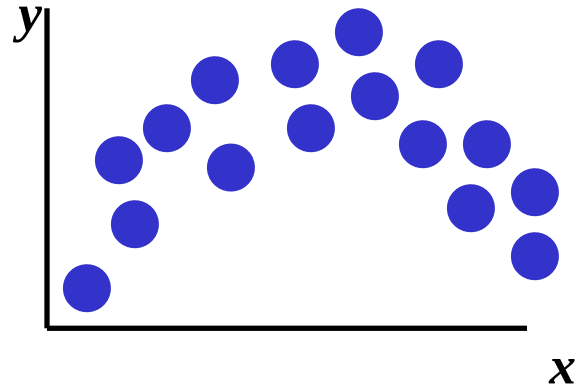
Испитување на графичкиот приказ

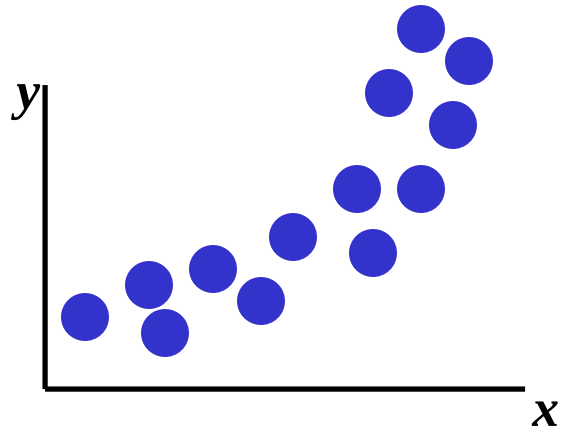
Линеарна врска





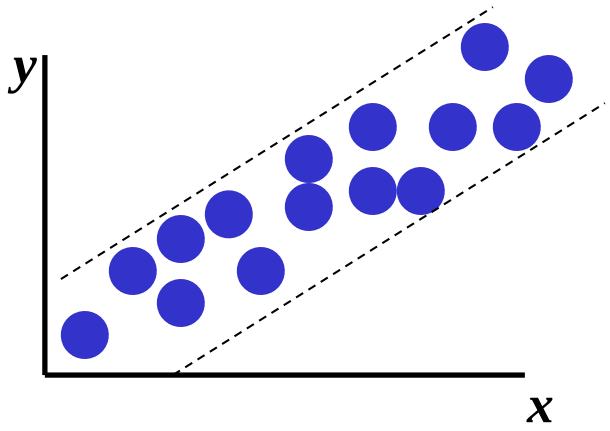
Криволиниска врска

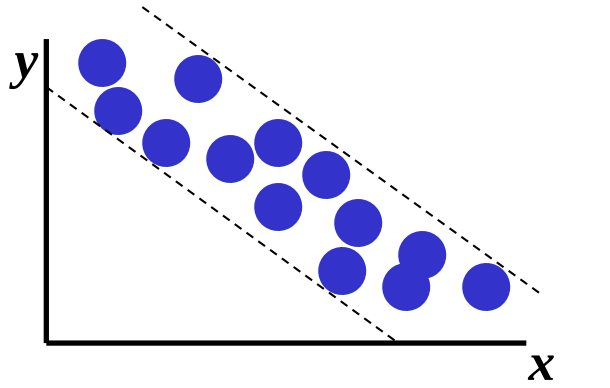




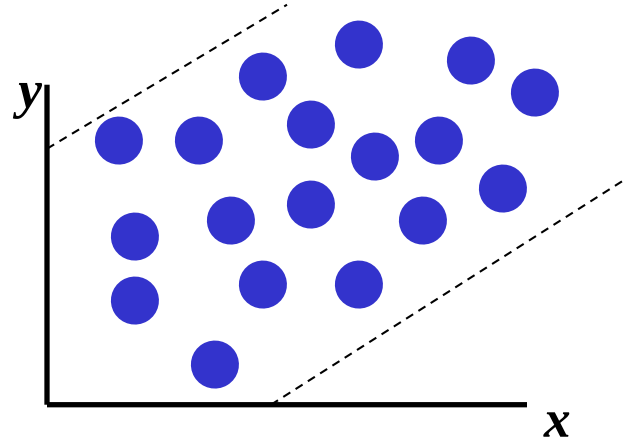
Испитување на графичкиот приказ

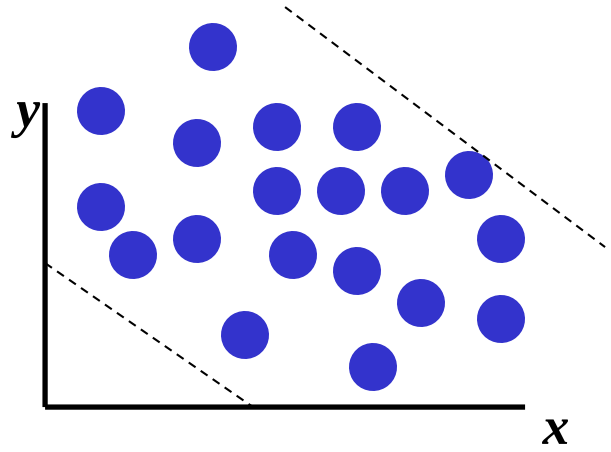
Силна поврзаност





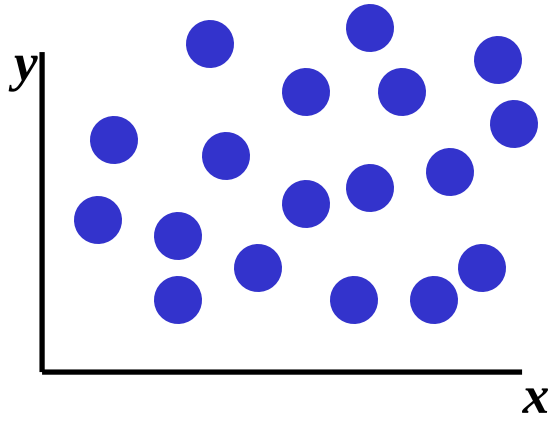
Слаба поврзаност

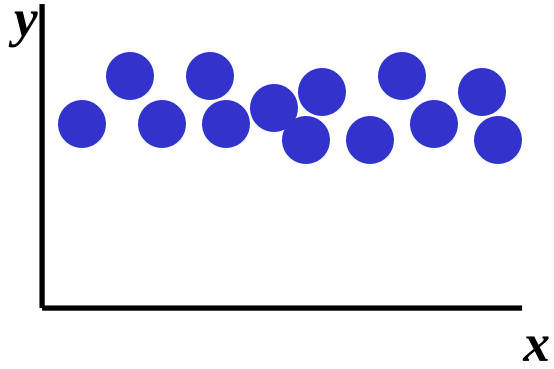




Испитување на графичкиот приказ

Не постои поврзаност





Коефициент на корелација

- Следно во анализата е да се воведат некои нумерички мерки кои ќе ја подржат претпоставката од графичкиот приказ за линеарна зависноста. Таква мерка е **коефициентот на корелација**.
- Коефициентот на корелацијата го мери правецот и јачината на линеарната врска помеѓу две квантитативни променливи.
  - Обично се означува со r.
  - Да претпоставиме дека имаме мерења на две променливи X и Y за n индивидуи.
  - Нека мерењата за првата индивидуа ги означиме со (x_1,y_1), мерењата за втората индивидуа со (x_2,y_2) и така натаму.
  - Нека $\bar{x}, \bar{s}_X, \bar{y}, \bar{s}_Y$ се просекот и коригираната стандардна девијација за променливите X и Y, соодветно. Тогаш корелацијата помеѓу X и Y е дадена со

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\bar{s}_X}\right)\left(\frac{y_i - \bar{y}}{\bar{s}_Y}\right).$$

Алтернативно пресметување на коефициентот на корелација

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} = \frac{ss_{XY}}{\sqrt{ss_X ss_Y}}$$

каде

$$ss_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad ss_X = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad ss_Y = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Или

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\bar{s}_X}\right)\left(\frac{y_i - \bar{y}}{\bar{s}_Y}\right) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\bar{s}_X \bar{s}_Y} = \frac{s_{XY}}{\bar{s}_X \bar{s}_Y},$$

каде

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad \bar{s}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \bar{s}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Коефициент на корелација

- Коефициентот на корелација не прави разлика помеѓу променливите, која е зависна, а која независна.
- Коефициентот на корелацијата се пресметува само ако и двете променливи се квантитативни.
- Бидејќи за пресметување на *r* се користат стандардизираните вредности на податоците, неговата вредност не се менува со промена на единицата мерка на било која променлива.
- Позитивно *r* укажува на позитивна асоцијација помеѓу променливите, а негативно *r* укажува на негативна асоцијација.

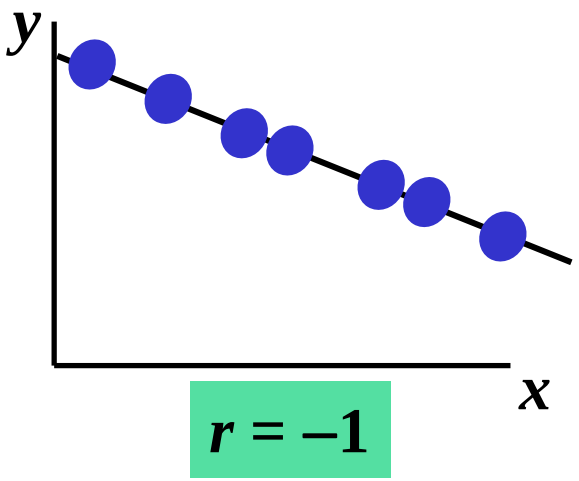
Коефициент на корелација

- Коефициентот на корелацијата r е секогаш број помеѓу –1 и 1. Вредностите на r блиску до 0 зборуваат за слаба линеарна поврзаност. Јачината на линеарната врска расте со растење на r кон 1 или опаѓање кон –1. Екстремните вредности $r = –1$ и $r = 1$ се јавуваат само во случај кога сите точки лежат на една права.

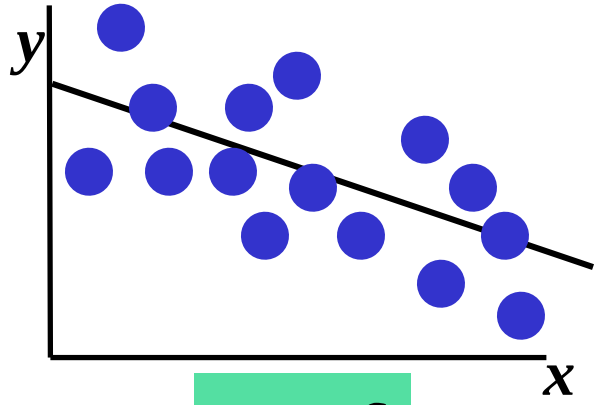
- **Коефициентот на корелацијата ја мери само јачината на линеарната врска помеѓу променливите.** Овој коефициент не ја одразува врската во вид на некоја друга крива линија, иако истата може да биде многу силна.

- Како и просекот и стандардната девијација, корелацијата не е резистентна, односно појава на значајни отстапувања ја менува корелацијата и затоа треба да се внимава.

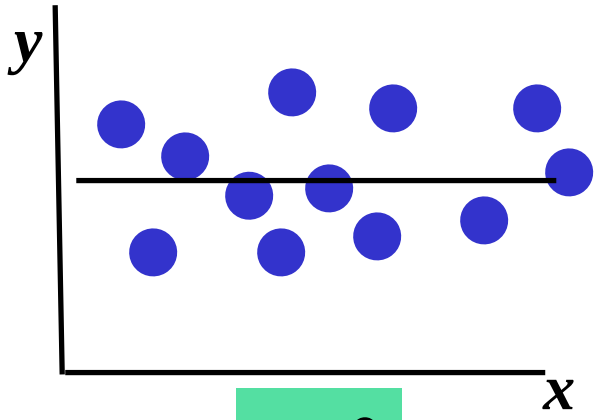
Примери за различни вредности на *r*



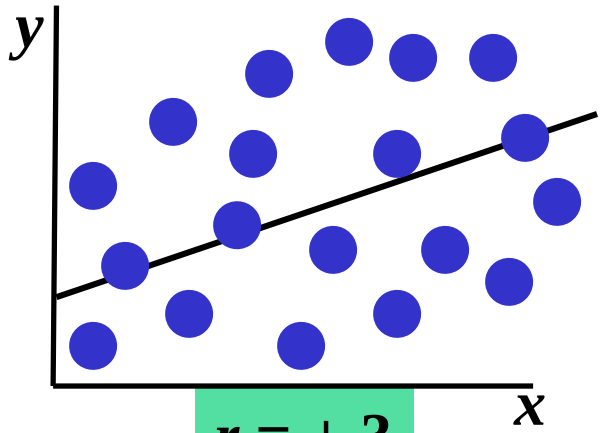
$r = -1$



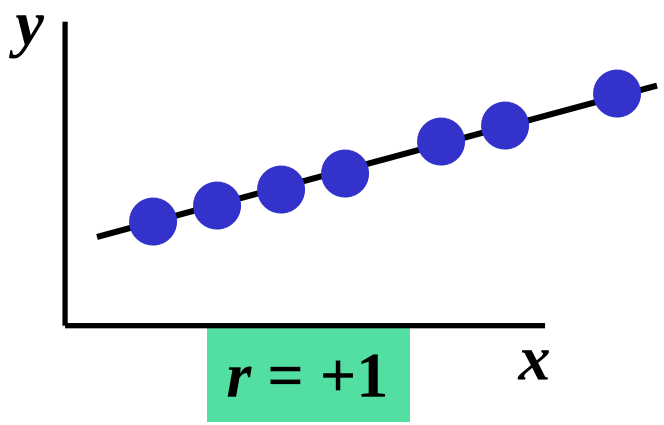
$r = -.6$



$r = 0$



$r = +.3$



$r = +1$

Проста линеарна регресија

- Нека се дадени парови на податоци (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), добиени со набљудување.

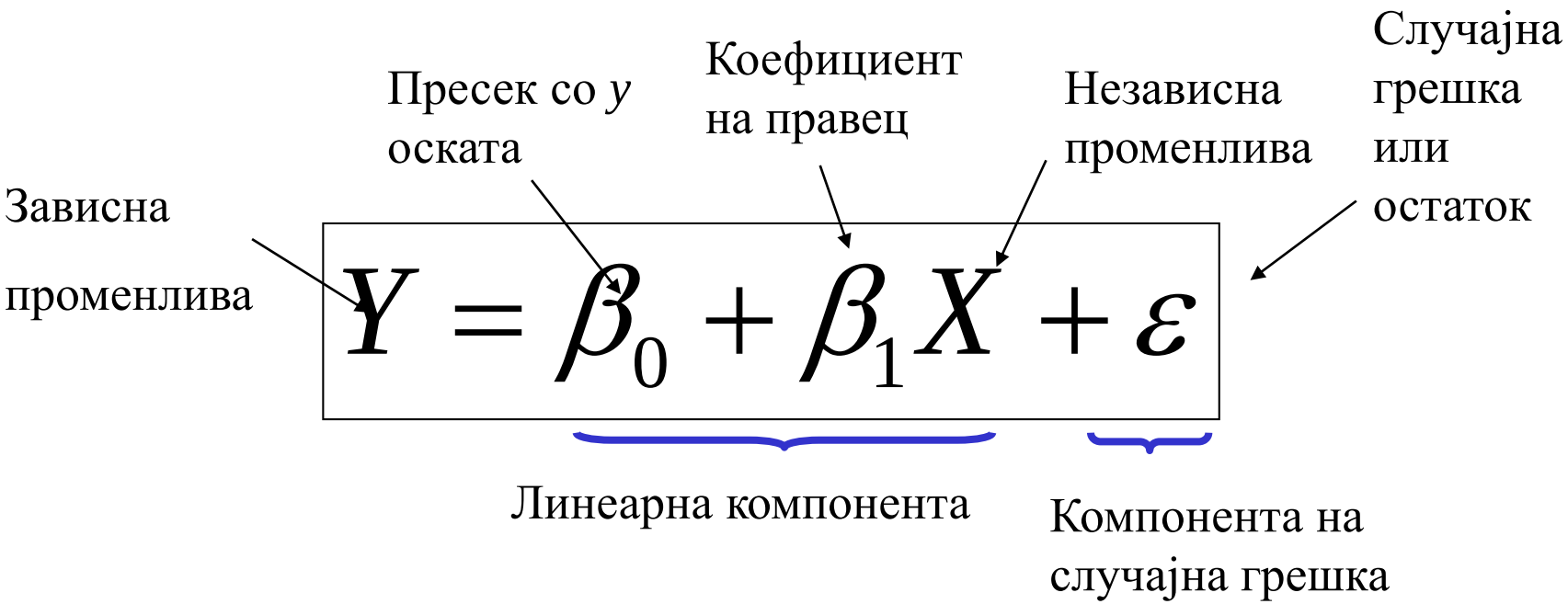
- Доколку се претпостави линеарна (проста) регресиона врска помеѓу координатите на точките, тогаш регресиониот модел е даден со

$$E(Y|x) = \beta_0 + \beta_1 x$$

или

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Линеарна регресија



Претпоставки за моделот

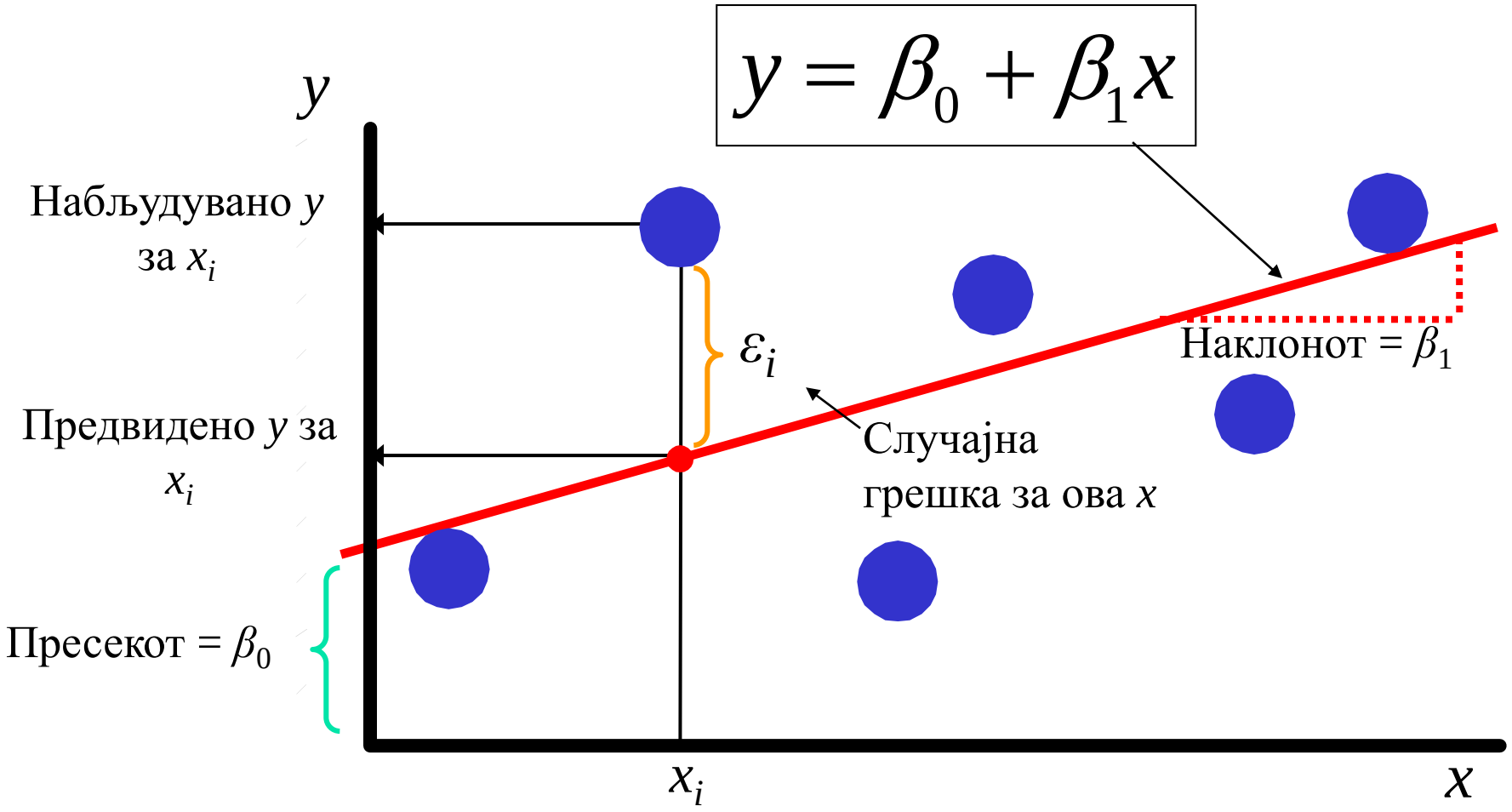
- *Претпоставка* 1. Математичкото очекување на ε е 0.

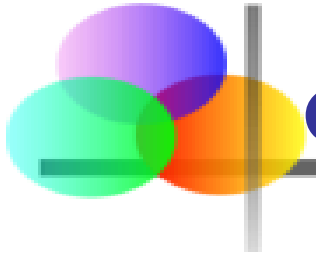
- *Претпоставка* 2. За сите вредности на независната променлива X, дисперзијата на ε е константна, σ^2 ;

- *Претпоставка* 3. Случајната грешка ε има нормална распределба, односно $\varepsilon \sim N(0, \sigma^2)$;

- *Претпоставка* 4. Случајните грешки се независни (во веројатносна смисла).

Модел на линеарна регресија



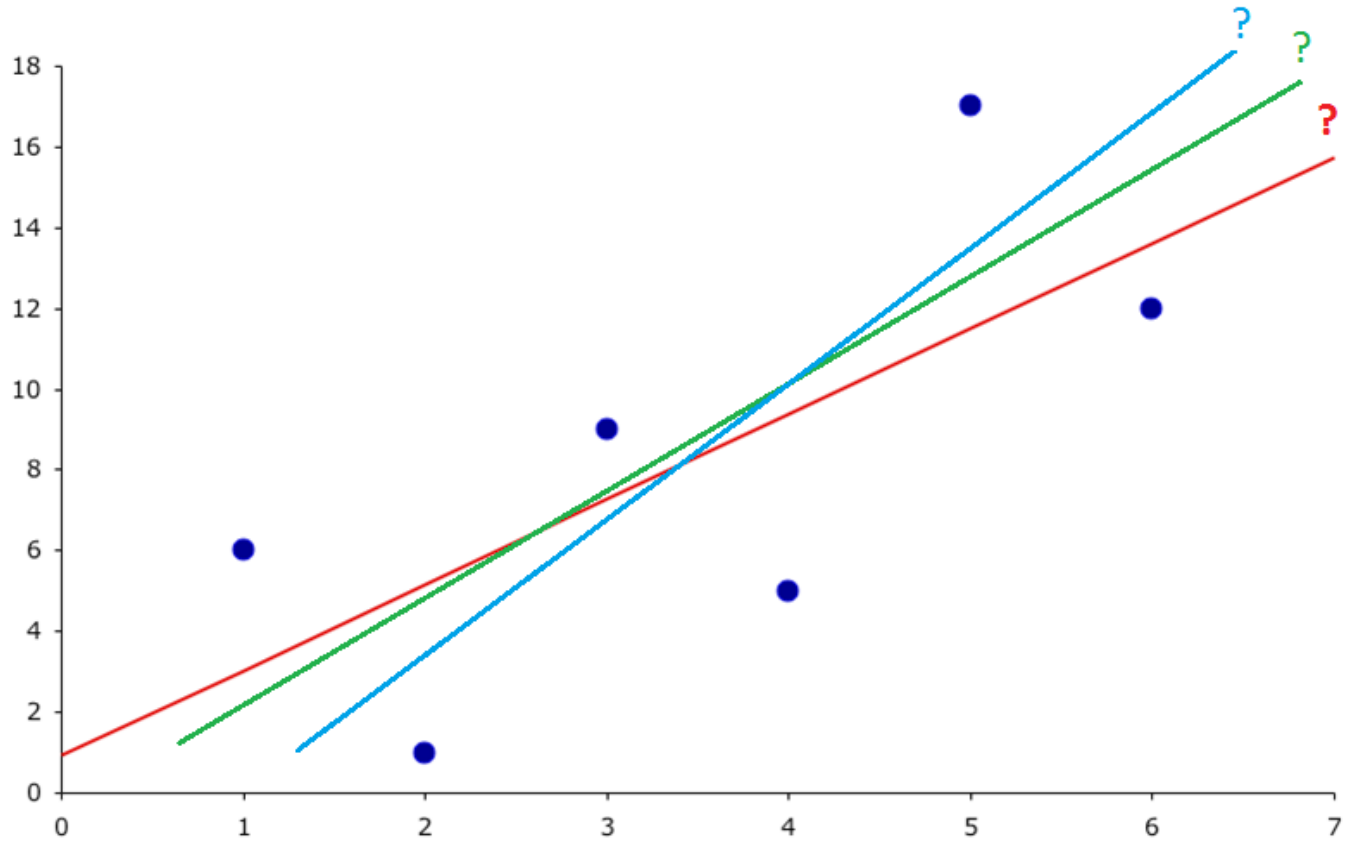


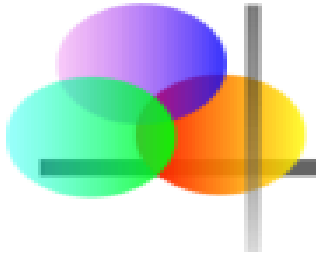
Оценување на параметрите на моделот

- За да го одбереме најсоодветниот модел за множество податоци, мораме да ги определиме непознатите параметри β_0, β_1, на моделот така што ќе се добие права која "најдобро одговара" на множеството набљудувани вредности (x_i,y_i), за $i = 1,\ldots, n$.

- Оценката на овие параметри се прави со метод на најмали квадрати.

Која е најсоодветната права ?





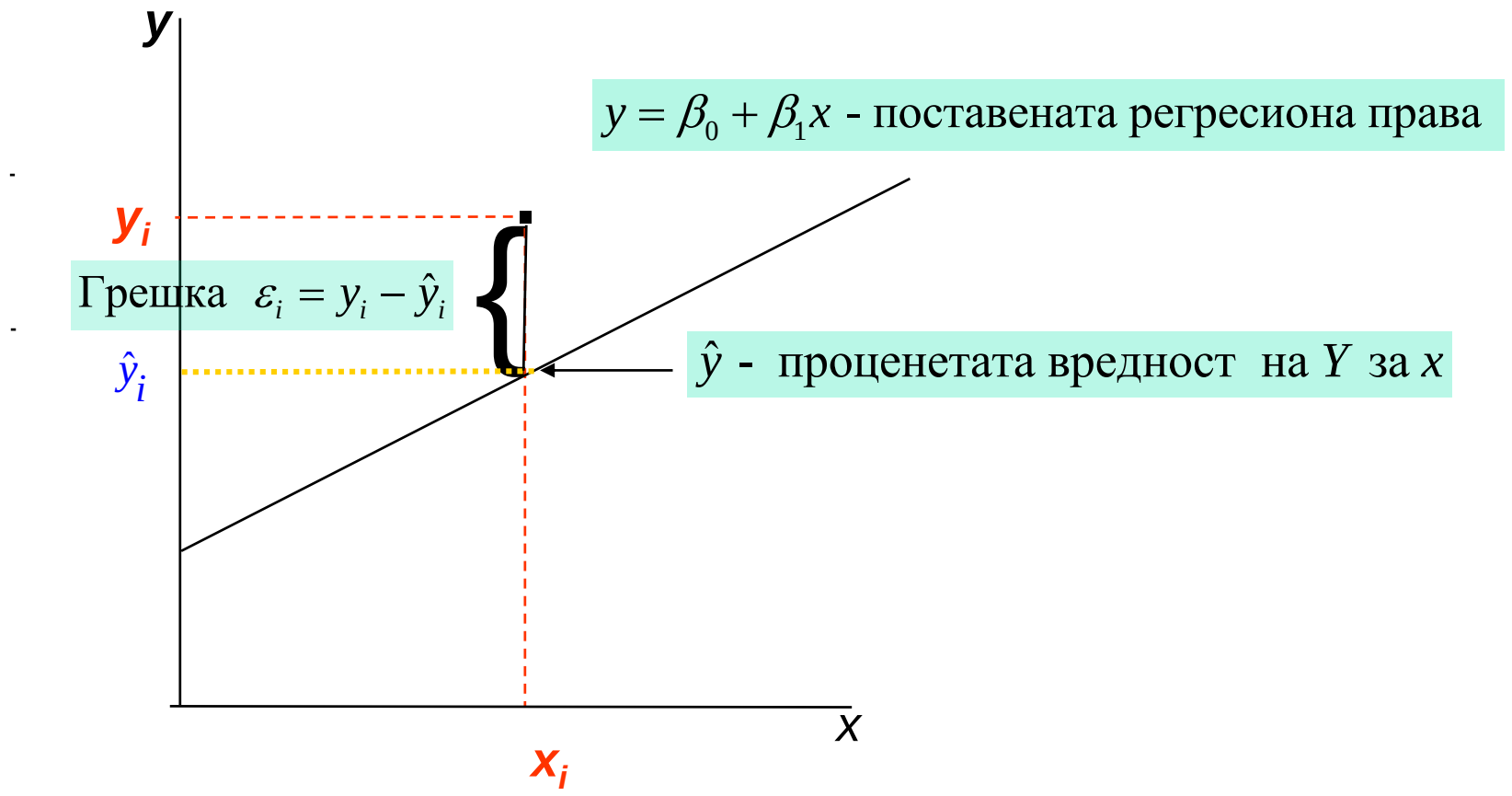
Метод на најмали квадрати

- **Регресиона права по метод на најмали квадрати** на *Y* по *x* е права која го прави збирот на квадратите на вертикалните растојанија на точките од податоците од дијаграмот до претпоставената права што е можно помал.

- Една од причините за популарноста на овој метод е тоа што поставувањето на права на овој начин има многу лесно и едноставно решение.

Метод на најмали квадрати

- Вертикалното растојание помеѓу набљудуваната вредност y_i на Y, и претпоставената (проценетата) вредност $\hat{y}_i = y(x_i) = \beta_0 + \beta_1 x_i$ е токму ε_i.

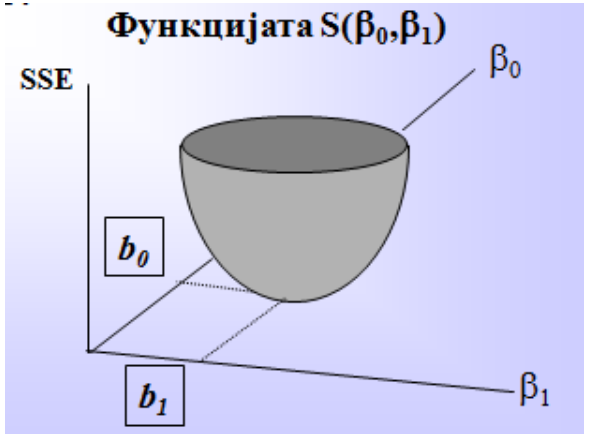


Метод на најмали квадрати

- Значи, за да ја определиме правата на регресија по метод на најмали квадрати треба врз база на примерокот да најдеме оценувачи за β_0 и β_1 за кои сумата

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

ќе прими најмала можна вредност.



Метод на најмали квадрати

- Се покажува дека функцијата $S(\beta_0, \beta_1)$ има минимум, ако

$$\hat{\beta}_0 = \bar{y} - \frac{ss_{xy}}{ss_x}\bar{x} \quad \text{и} \quad \hat{\beta}_1 = \frac{ss_{xy}}{ss_x}.$$

- Тогаш проценетата права гласи

$$\hat{y} = \beta_0 + \beta_1 x$$

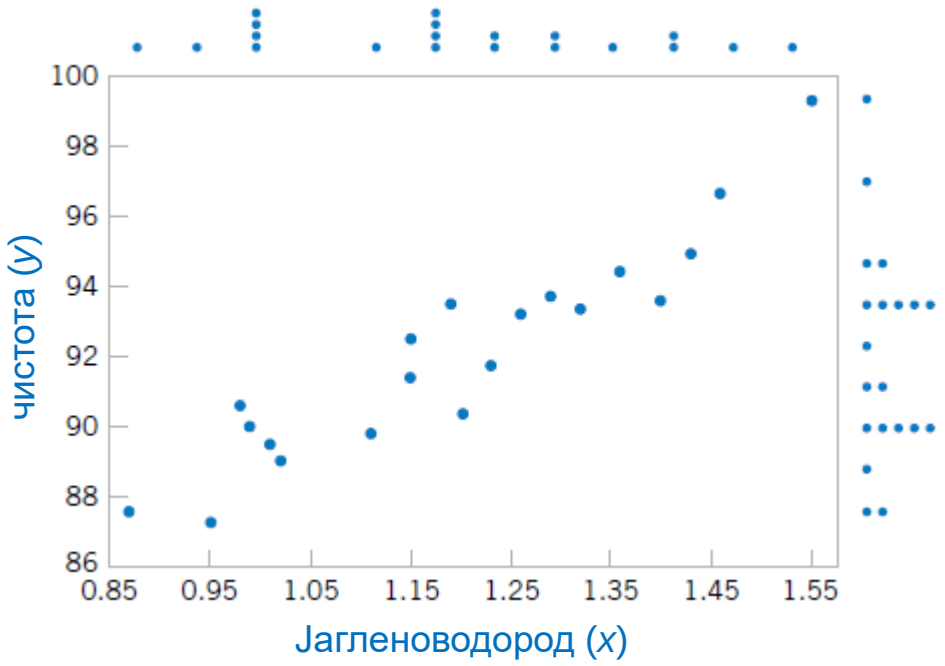
и истата може да се користи за предвидување на вредности за обележјето Y за дадени вредности на обележјето X.

- Правата на регресија може да се користи за предвидување на вредности на y само за вредности на x во распонот на набљудуваните, или многу блиску до најмалата или најголемата вредност на x.

Пример 1

- Да се определи правата на регресија која ја изразува зависноста на чистотата (y) на кислород произведен со хемиска дестилација од процентот (x) на јагленоводород што е присутен во главниот кондензатор на единицата за дестилација. Направени се 20 мерења и резултатите се претставени во табелата.

бр. на набљудување	јагленоводород x (%)	чистота y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33



Решение на Пример 1

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
0.99	90.01	89.1099	0.98	8101.80
1.02	89.05	90.831	1.04	7929.90
1.15	91.43	105.145	1.32	8359.44
1.29	93.74	120.925	1.66	8787.19
1.46	96.73	141.226	2.13	9356.69
1.36	94.45	128.452	1.85	8920.80
0.87	87.59	76.2033	0.76	7672.01
1.23	91.77	112.877	1.51	8421.73
1.55	99.42	154.101	2.40	9884.34
1.40	93.65	131.11	1.96	8770.32
1.19	93.54	111.313	1.42	8749.73
1.15	92.52	106.398	1.32	8559.95
0.98	90.56	88.7488	0.96	8201.11
1.01	89.54	90.4354	1.02	8017.41
1.11	89.85	99.7335	1.23	8073.02
1.20	90.39	108.468	1.44	8170.35
1.26	93.25	117.495	1.59	8695.56
1.32	93.41	123.301	1.74	8725.43
1.43	94.98	135.821	2.04	9021.20
0.95	87.33	82.9635	0.90	7626.53
23.92	1843.21	2214.66	29.29	170044.53

$$\bar{x} = \frac{23.92}{20} = 1.196$$

$$\bar{y} = \frac{1843.21}{20} = 92.1605$$

$$ss_X = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$= 29.29 - 20 \cdot 1.196^2 = 0.68$$

$$ss_{XY} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$= 2214.66 - 20 \cdot 1.196 \cdot 92.1605 = 10.18$$

Решение на Пример 1 - продолжение

$$\bar{x} = 1.196 \qquad \bar{y} = 92.1605$$

$$ss_X = 0.68 \qquad ss_{XY} = 10.18$$

- За коефициентите на правата на регресија, се добива:

$$\hat{\beta}_1 = \frac{ss_{XY}}{ss_X} = \frac{10.18}{0.68} = 14.97$$

$$\hat{\beta}_0 = \bar{y} - \frac{ss_{XY}}{ss_X}\bar{x} = 92.1605 - 14.97 \cdot 1.196 = 74.25$$

- Оттука, бараната права на регресија е

$$y = 74.25 + 14.97x.$$

- Ако треба да се оцени вредноста на y за $x = 1.5$, тогаш се добива:

$$y(1.5) = 74.25 + 14.97 \cdot 1.5 = 96.705.$$

Алтернативен облик на права на регресија

- Со едноставни алгебарски трансформации, правата на регресија може да се запише и во следниот облик:

$$y - \bar{y} = r \frac{\bar{s}_Y}{\bar{s}_X}(x - \bar{x}).$$

Пример 2

Случајно се избрани 10 студенти и се прашани колку часа го подготвувале испитот по Бизнис статистика. Нивните одговори се споредени со поените кои ги освоиле на испит. Максималниот број на поени кои може да се освојат на испит е 100.

Часови за подготовка (X)	12	31	22	7	10.8	25	15.6	23.5	17.2	14
Поени (Y)	45	60	88	25	42	85	51	80	60	53

Да се определи равенката на правата на регресија на бодовите (y) во однос на времето за подготовка на испитот, изразено во часови (x). Што може да се каже за коефициентот на корелација? Колку е проценетиот број на поени за студент кој испитот го спремал 20 часа?

Решение на Пример 2

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
12	45	540	144	2025
31	60	1860	961	3600
22	88	1936	484	7744
7	25	175	49	625
10.8	42	453.6	117	1764
25	85	2125	625	7225
15.6	51	795.6	243	2601
23.5	80	1880	552	6400
17.2	60	1032	296	3600
14	53	742	196	2809
178	589	11539.2	3667	38393

$$\bar{x} = \frac{178}{10} = 17.8 \quad \bar{y} = \frac{589}{10} = 58.9$$

$$\bar{s}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{10}\cdot 3667 - 17.8^2 = 49.86, \quad \bar{s}_X = \sqrt{\bar{s}_X^2} = 7.06$$

$$\bar{s}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{1}{10}\cdot 38393 - 58.9^2 = 370.09, \quad \bar{s}_Y = \sqrt{\bar{s}_Y^2} = 19.24$$

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = \frac{1}{10}\cdot 11539.2 - 17.8\cdot 58.9 = 105.5$$

$$r = \frac{s_{XY}}{\bar{s}_X \bar{s}_Y} = \frac{105.5}{7.06\cdot 19.24} = 0.77$$

Од добиената вредност *r* може да се заклучи дека постои значителна линеарна зависност помеѓу овие две променливи.

Решение на Пример 2 продолжение

- Сега, за правата на регресија на y во однос на x, се добива:

$$y - \bar{y} = r\frac{\bar{s}_Y}{\bar{s}_X}(x - \bar{x})$$

$$y - 58.9 = 0.77\,\frac{19.24}{7.06}(x - 17.8)$$

$$y = 2.1x + 21.55$$

- Овде, коефициентот на корелација е позитивен, исто како и коефициентот β_1, затоа може да се заклучи дека бројот на поени ќе се зголемува со зголемување на времето за подготовка на испитот.
- Предвидениот (проценетиот) број на поени за студент кој учел 20 часа е

$$y(20) = 2.1 \cdot 20 + 21.55 = 63.55$$

Права на регресија на X по Y

- Во некои случаи, потребно е X и Y да си ги заменат местата: Y да биде независна, а X да биде зависна променлива. Во овој случај, x и y, како и карактеристиките на X на Y си ги менуваат местата, па равенката добива облик:

$$x - \bar{x} = r \frac{\bar{s}_X}{\bar{s}_Y} (y - \bar{y}).$$

Пример 3

За неколку случајно избрани семејства добиени се податоци за дневната потрошувачка на млеко (во литри) и бројот на членови на семејството

Број на членови (*X*)	**2**	**4**	**3**	**6**	**3**	**4**	**3**	**4**
Потрошувачка на млеко (*Y*)	1	3	1	4	2	2	2	3

Да се определат двете прави на регресија. Потоа да се процени потрошувачката на млеко во петчлено семејство.

Решение на Пример 3

x_i	y_i	x_iy_i	x_i^2	y_i^2
2	1	2	4	1
4	3	12	16	9
3	1	3	9	1
6	4	24	36	16
3	2	6	9	4
4	2	8	16	4
3	2	6	9	4
4	3	12	16	9
29	**18**	**73**	**115**	**48**

$$\bar{x} = \frac{29}{8} = 3.625 \quad \bar{y} = \frac{18}{8} = 2.25$$

$$\bar{s}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2 = \frac{1}{8}\cdot 115 - 3.625^2 = 1.234, \quad \bar{s}_X = \sqrt{\bar{s}_X^2} = 1.11$$

$$\bar{s}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}y_i^2 - \bar{y}^2 = \frac{1}{8}\cdot 48 - 2.25^2 = 0.935, \quad \bar{s}_Y = \sqrt{\bar{s}_Y^2} = 0.967$$

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \bar{x}\bar{y} = \frac{1}{8}\cdot 73 - 3.625\cdot 2.25 = 0.96875$$

$$r = \frac{s_{XY}}{\bar{s}_X\,\bar{s}_Y} = \frac{0.96875}{1.11\cdot 0.967} = 0.9025$$

Бидејќи r е блиску до 1 следува дека постои силна линеарна зависност.

Решение на Пример 3 - продолжение

- Правата на регресија на Y по X добива облик:

$$y - \bar{y} = r\frac{\bar{s}_Y}{\bar{s}_X}(x - \bar{x})$$

$$y - 2.25 = 0.9025 \cdot \frac{0.967}{1.11}(x - 3.625)$$

$$y = -0.6 + 0.786x$$

- Втората права на регресија на X по Y ја добиваме со замена на местата на x и y:

$$x - \bar{x} = r\frac{\bar{s}_X}{\bar{s}_Y}(y - \bar{y})$$

$$x - 3.625 = 0.9025 \cdot \frac{1.11}{0.96}(y - 2.25)$$

$$x = 1.277 + 1.044y$$

Потрошувачката во петчлено семејство може да се оцени ако се замени $x = 5$:
$y(5) = -0.6 + 0.786 \cdot 5 = 3.33$

Интерпретација на резултатите

- Ако y_i е набљудуваната вредност на променливата Y, а $\bar{y}$ е просечната вредност на набљудуваните y_i, $i = 1, \ldots, n$, разликата $y_i - \bar{y}$ се нарекува *вкупно отстапување* на податокот од просекот.

- Ако $\hat{y}_i$ е вредноста на податокот добиена од правата на регресија, разликата $\hat{y}_i - \bar{y}$ се нарекува *објаснето отстапување* (или отстапување што се должи на моделот) и покажува за колку се намалува вкупното отстапување кога ќе се постави регресионата права на податоците.

- На крајот, разликата $y_i - \hat{y}_i$ се нарекува *необјаснето отстапување*, односно дел од вкупното варирање кој не е објаснет со воведувањето на регресионата права.

Објаснето и необјаснето отстапување

- *SST* = вкупна сума на квадрати

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

  - Го мери варирањето на y_i околу нивниот просек

- *SSE* = сума на квадрати на грешки

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

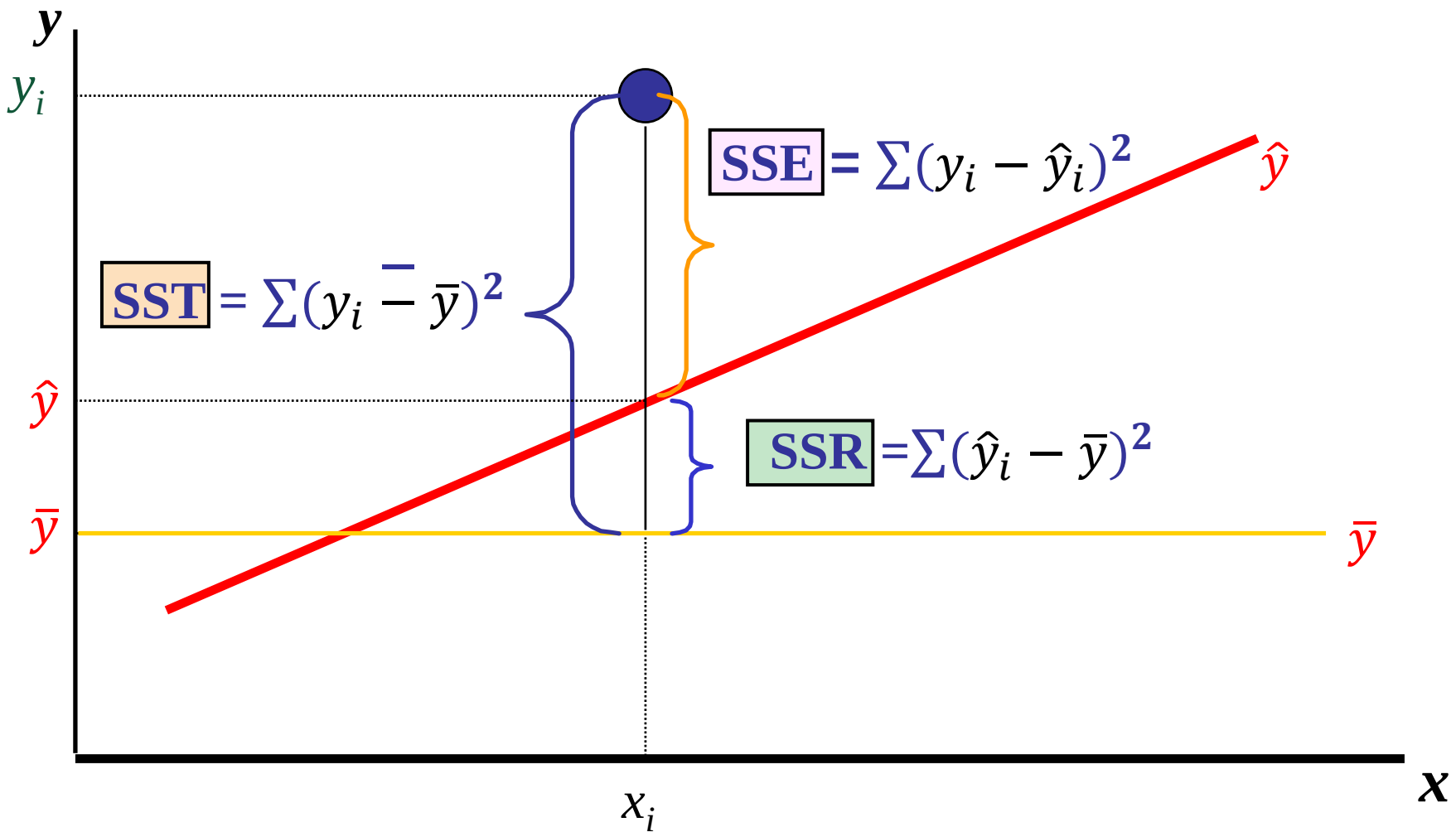
  - Варирање што се должи на други причини надвор од релацијата меѓу *x* и *y*

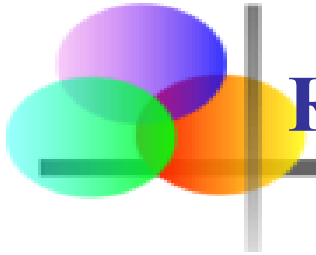
- *SSR* = сума на квадрати на регресија

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

  - Објаснето варирање што се должи на линеарната врска на *x* и *y*

Објаснето и необјаснето отстапување



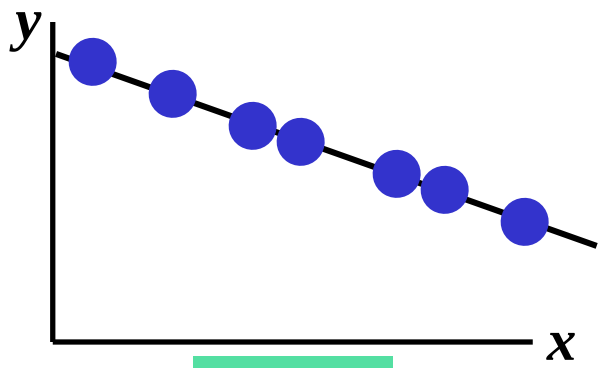


Коефициент на детерминираност

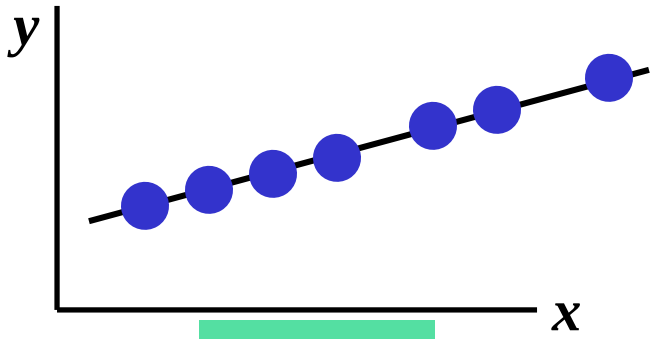
- Бројот R^2 = SSR/SST се нарекува *коефициент на детерминираност* и ја мери јачината на совпаѓањето на правата на регресија со податоците.

- Овој број има вредност помеѓу 0 и 1 и колку е поблиску до 1 толку е подобро совпаѓањето, односно толку подобар е регресиониот модел.

Примери за вредности на R^2



$R^2 = 1$



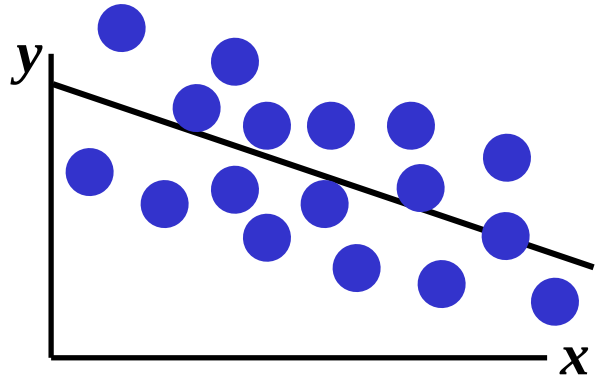
$R^2 = +1$

$R^2 = 1$

Перфектна линеарна асоцијација меѓу x и y.

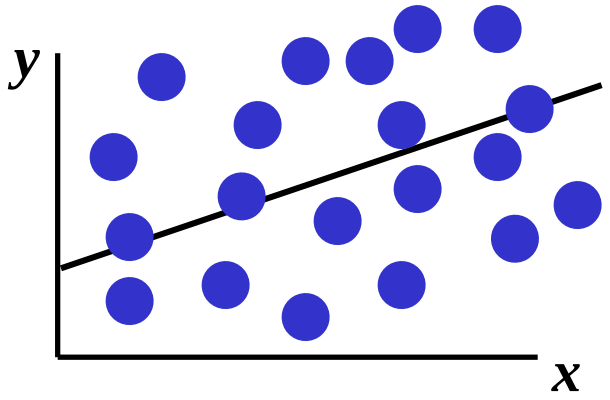
100% од варирањето на y е објаснето со варирањето на x вредностите.

Примери за вредности на R^2



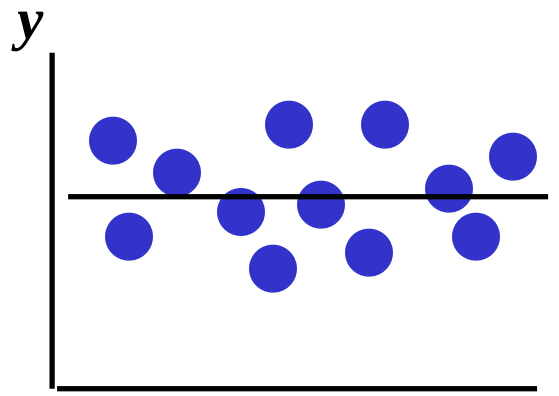
$0 < R^2 < 1$

Послаба линеарна поврзаност на x и y.



Дел од варирањето во y е објаснето со варирањето во x.

Примери за вредности на R^2



$R^2 = 0$

$R^2 = 0$

Не постои линеарна поврзаност меѓу x и y.

Вредноста на Y не зависи од x.