Тема 12.
Корреляция и регрессия

12.1. Корреляция

12.2. Значимость коэффициента корреляции

12.3. Регрессия

12.4. Надежность прогноза

Примеры

1. **Менеджер** интересуется, зависит ли объем продаж в этом месяце от объема рекламы в этом же периоде?
2. **Преподаватель** хочет выяснить, есть ли зависимость между количеством часов, потраченных студентом на занятия, и результатами экзамена?
3. **Врач** исследует, влияет ли кофеин на сердечные болезни и существует ли связь между возрастом человека и его кровяным давлением?
4. **Зоолог** стремится узнать, есть ли связь между весом определенного животного при рождении и его продолжительностью жизни.
5. **Социолог** исследует, какова связь между уровнем преступности и уровнем безработицы в регионе? Есть ли зависимость между расходами на жилье и совокупным доходом семьи? Связаны ли доход от профессиональной деятельности и продолжительность образования?

На эти вопросы можно ответить, используя методы корреляционного и регрессионного анализа, рассмотренные в материалах этой лекции.

Постановка проблемы

Наша цель – научиться отвечать на четыре вопроса:

Вопрос 1. **Существует ли связь** между двумя или более переменными?

Вопрос 2. Какой **тип** имеет эта связь?

Вопрос 3. Насколько она **сильна**?

Вопрос 4. Какой можно сделать **прогноз**, основываясь на этой связи?

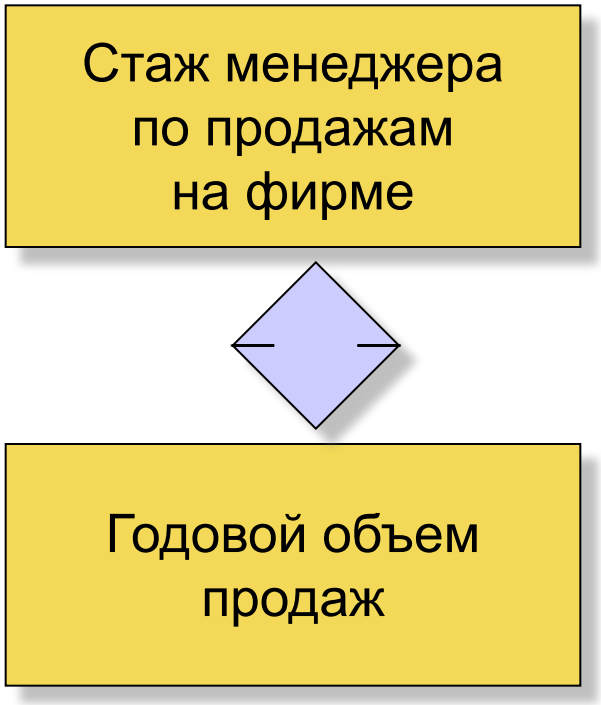
Методы

Корреляция – статистический метод, позволяющий определить, существует ли зависимость между переменными и на сколько она сильна.

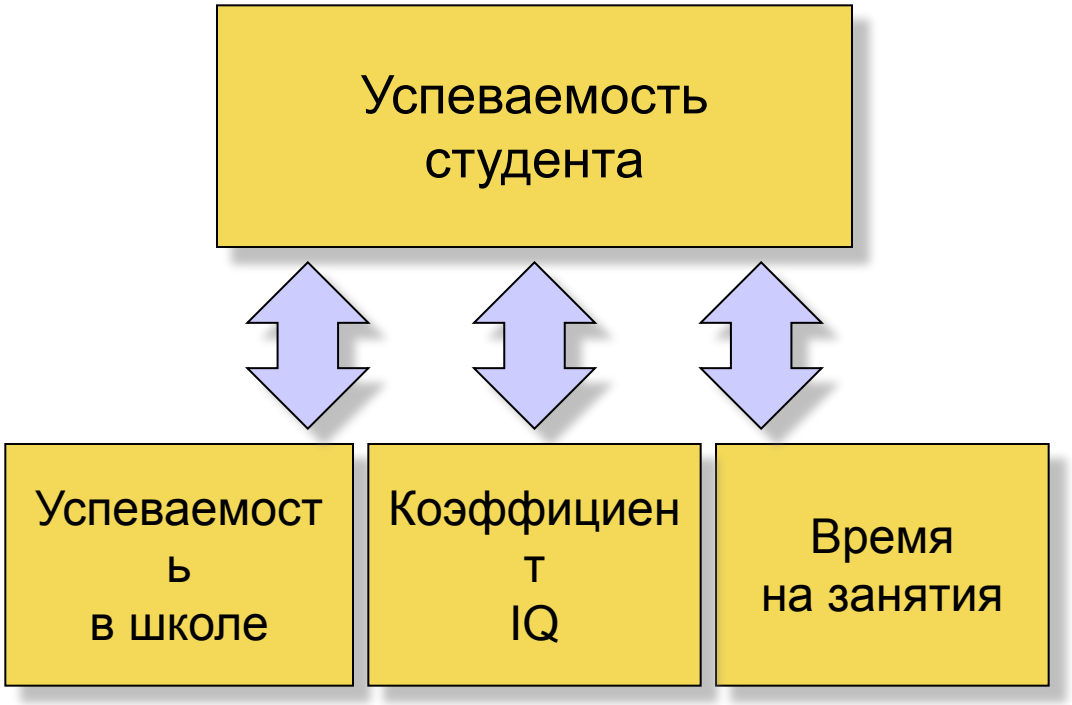
Регрессия – статистический метод, который используется для описания характера связи между переменными (положительная или отрицательная, линейная или нелинейная зависимость).

Простая и множественная связь

Простая связь означает изучение двух переменных.



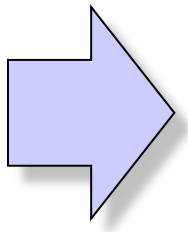
Множественная связь означает изучение несколько переменных.

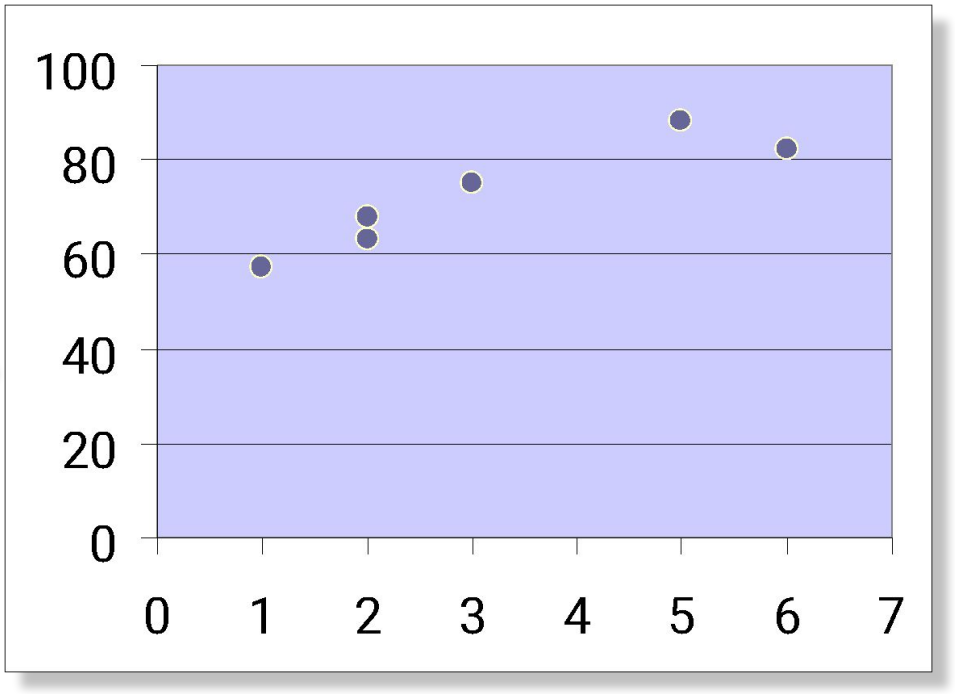


Визуальный анализ связи

Рассматриваем две переменные: «продолжительность занятий» студентов перед экзаменом и «итоговая оценка» (из 100 балов). Пытаемся визуально определить связь. Правда ли, что **чем меньше времени занятий, тем выше оценка?**

Студент	Часы x	Оценка y
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75





Независимая и зависимая переменные

Независимая переменная – это та переменная в регрессии, которую можно изменять. В данном случае, переменная «количество часов занятий» является независимой и обозначается как переменная *x*.

Зависимая переменная – это переменная в регрессии, которую нельзя изменять. «Экзаменационная оценка» является зависимой переменной. Она обозначается *y*.

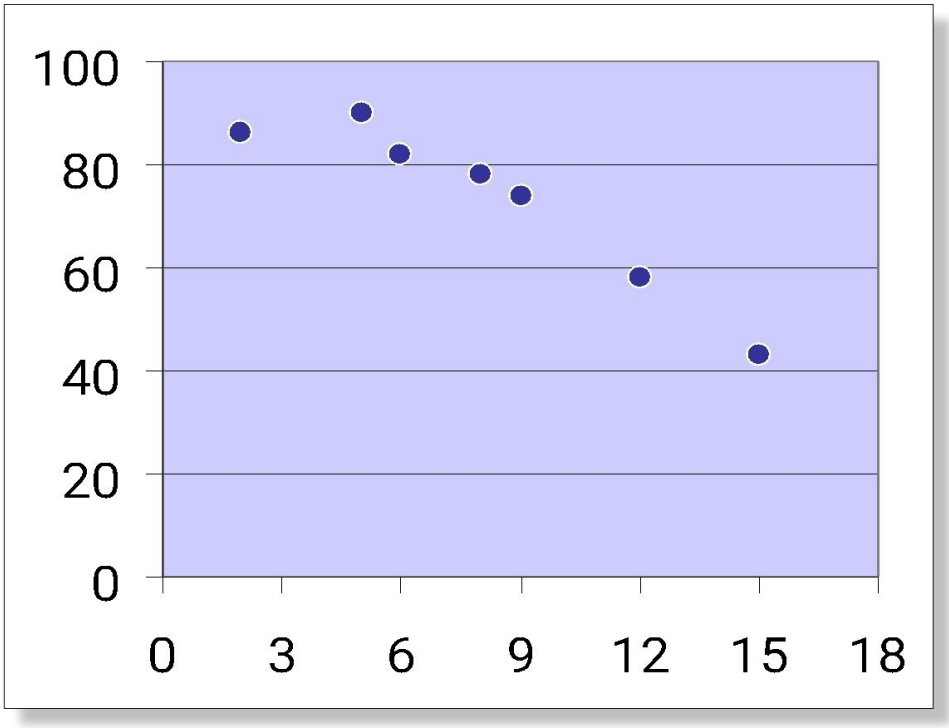
Причиной такого разделения переменных является то, что *предполагается*, что оценка, которую получает студент, зависит от количества часов, которые он посвятил занятиям. Предполагается также, что студенты могут регулировать количество часов, которое они тратят на занятия.

Не всегда можно ясно определить, какая переменная зависимая, а какая независимая, и выбор иногда делается произвольно.

Положительная и отрицательная зависимость

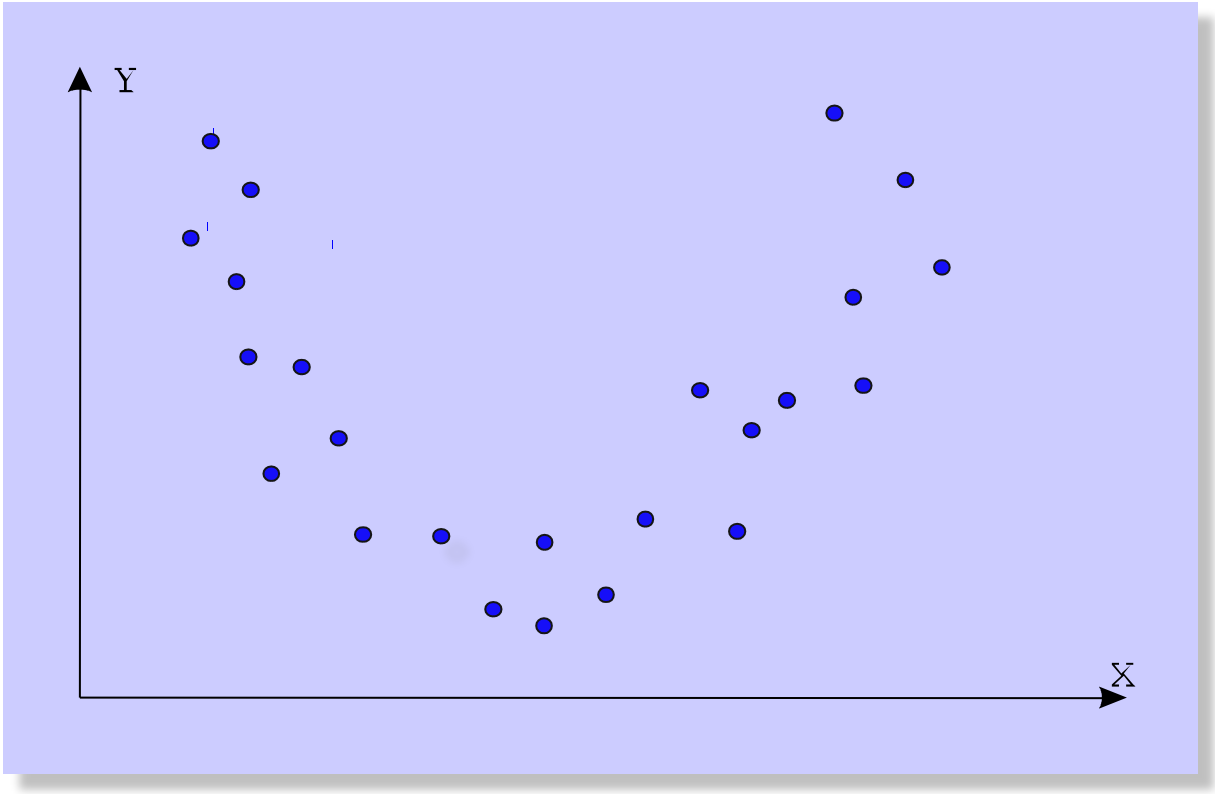
Визуально видно, что имеет место линейная зависимость, которая отрицательна. Это означает, что увеличение переменной *x* приводит к уменьшению второй переменной *y*.

Студент	Пропущено *x*	Оценка *y*
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78



Нелинейная зависимость

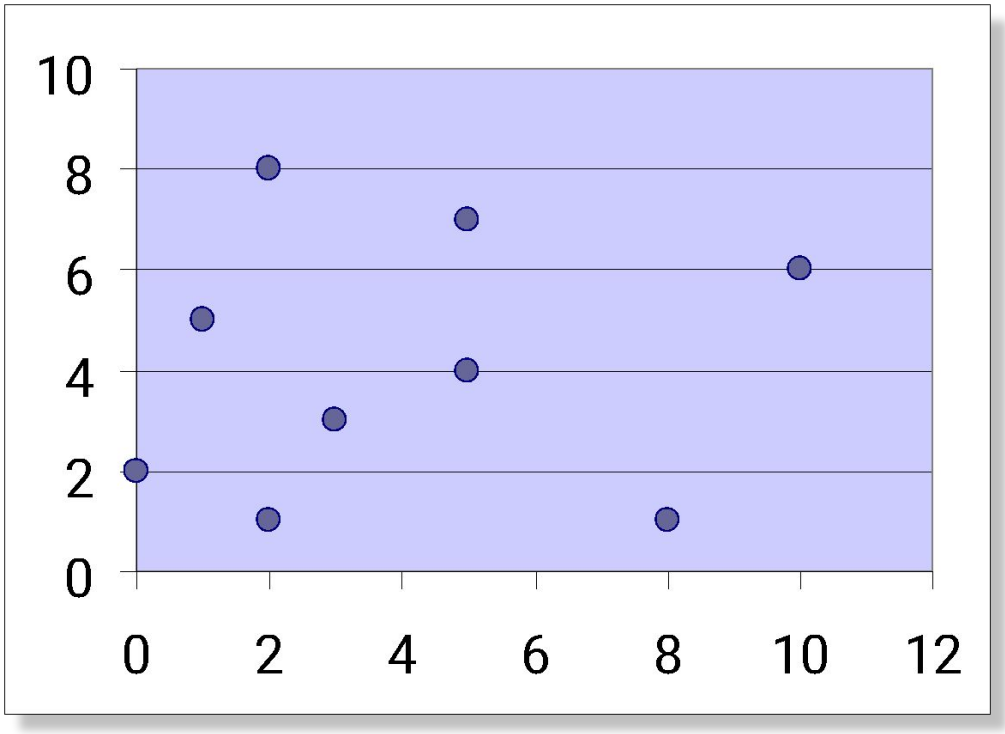
График показывает, что имеется зависимость, которая не является линейной. Возможно, эта зависимость квадратичная или какая-то иная.



Отсутствие зависимости

Студент	Часы занятий *x*	Бутылки пива *y*
A	3	3
B	0	2
C	2	1
D	5	7
E	8	1
F	5	4
G	10	6
H	2	8
I	1	5

График сообщает нам об отсутствии зависимости продолжительности занятий в неделю от количества выпиваемого пива (в бутылках).



12.1. Корреляция

Связь между двумя переменными

Коэффициент корреляции

Коэффициент корреляции измеряет силу и направление связи между двумя переменными.

Обозначения:

Выборочный коэффициент корреляции r

Коэффициент корреляции генеральной совокупности ρ

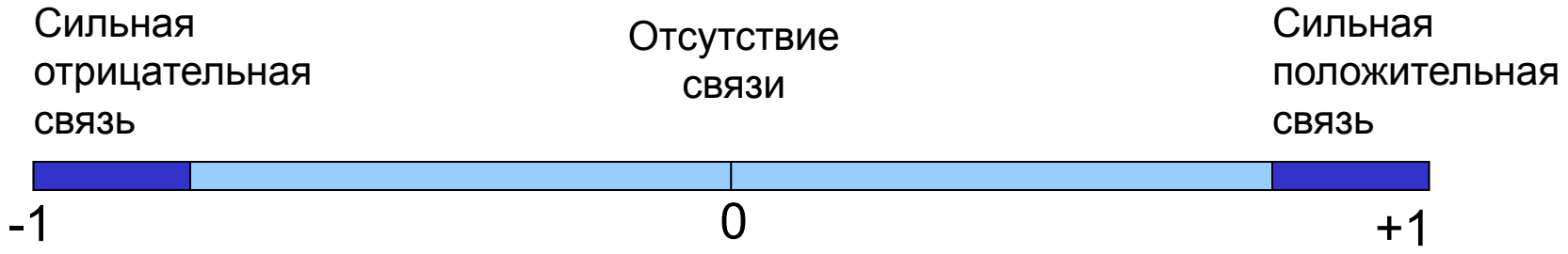
Значения коэффициента корреляции

Коэффициент корреляции изменяется на отрезке от –1 до +1.

Если между переменными существует сильная положительная связь, то значение *r* будет близко к +1.

Если между переменными существует сильная отрицательная связь, то значение *r* будет близко к –1.

Когда между переменными нет линейной связи или она очень слабая, значение r будет близко к 0.



Формула для вычисления *r*

Коэффициент корреляции вычисляется по формуле:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Это, так называемый, коэффициент корреляции Пирсона, равный произведению моментов. Он назван по имени статистика Карла Пирсона, который первый провел исследования в этой области.

Вторая формула для вычисления *r*

После несложных преобразований, из первой формулы можно получить другую формулу для коэффициента.

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\left(\sum x^2\right) - \left(\sum x\right)^2\right] \cdot \left[n\left(\sum y^2\right) - \left(\sum y\right)^2\right]}}$$

Как мы увидим, она более пригодна для вычисления коэффициента при помощи таблиц.

Пример вычисления

Вычислим коэффициент корреляции для примера со студентами.

Студент	Часы x	Оценка y
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

Шаг 1. Достроим таблицу

Достраиваем таблицу тремя столбцами и итоговой строкой. Проводим необходимые вычисления.

Студент	Часы x	Оценка y	xy	x^2	y^2
A	6	82	492	36	6724
B	2	63	126	4	3969
C	1	57	57	1	3249
D	5	88	440	25	7744
E	2	68	136	4	4624
F	3	75	225	9	5625
	$\Sigma x=19$	**$\Sigma y=433$**	**$\Sigma xy=1476$**	**$\Sigma x^2=79$**	**$\Sigma y^2=31935$**

Шаги 2-3. Подставим в формулу, получим ответ

Подставим данные в формулу и найдем *r :*

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{\left[n(\sum x^2) - (\sum x)^2\right] \cdot \left[n(\sum y^2) - (\sum y)^2\right]}} =$$

$$= \frac{6 \cdot 1476 - 19 \cdot 433}{\sqrt{\left[6 \cdot 79 - 19^2\right] \cdot \left[6 \cdot 31935 - 433^2\right]}} = 0{,}922$$

Ответ. Значение коэффициента корреляции равно 0,922. Это означает, что существует сильная положительная связь. Мы видели эту связь графически.

12.2. Значимость коэффициента корреляции

Проверка гипотезы

Постановка проблемы

Коэффициент корреляции генеральной совокупности ρ – это корреляция, вычисленная с использованием всевозможных пар значений признаков (x,y) генеральной совокупности.

Требуется

Оценить коэффициент корреляции генеральной совокупности ρ на основе значения коэффициента корреляции выборки r.

Условия

Выборочный коэффициент корреляции r используется для оценки ρ, если выполнены следующие предположения:

- Переменные x и y *линейно* зависимы
- Переменные являются *случайными*
- Обе переменные имеют *нормальное распределение*

Последовательность действий

Чтобы принять верное решение, воспользуемся процедурой проверки гипотезы. Она включает традиционные пять шагов:

Шаг 1. Сформулировать гипотезы.

Шаг 2. Построить критическую область.

Шаг 3. Вычислить значение критерия.

Шаг 4. Сравнить, принять решение.

Шаг 5. Написать ответ.

Гипотезы

Гипотезы сформулированы следующим образом.

Основная гипотеза H_0: $\rho = 0$
Альтернативная гипотеза H_1: $\rho \neq 0$

Основная гипотеза утверждает, что не существует корреляции между признаками *x* и *y* в генеральной совокупности. Альтернативная гипотеза утверждает, что корреляция между признаками *x* и *y* в генеральной совокупности значима.

Когда основная гипотеза отвергается на определенном уровне значимости, это значит, что существует значимое различие между значением *r* и 0. Когда основная гипотеза принимается, это значит, что значение *r* не сильно отличается от 0 и является случайным.

Статистика и критическая область

Для проверки гипотезы используется t-критерий с df = n – 2 степенями свободы:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Границы двусторонней критической области находятся при помощи таблиц значений t-распределения.

Пример

Задача. Рассчитан коэффициент корреляции и его значение оказалось равно 0,897. Выборка содержала 6 пар. На уровне значимости 0,05 проверить гипотезу о значимости коэффициента корреляции.

Решение.

Шаг 1. H_0: $\rho = 0$ H_1: $\rho \neq 0$

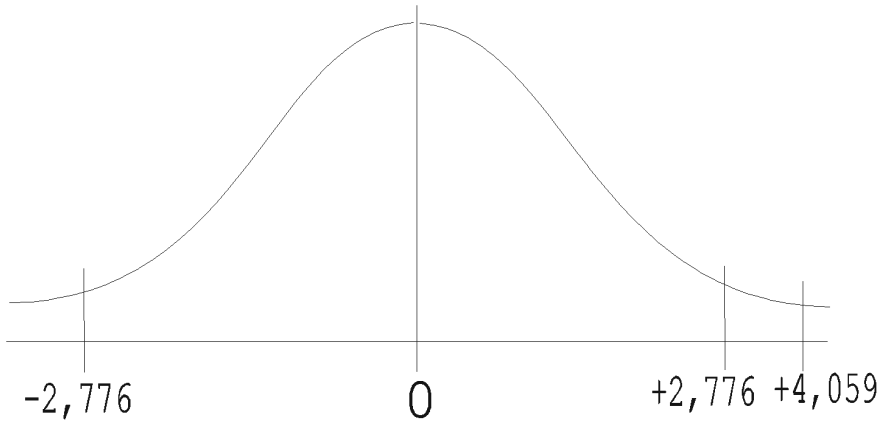
Шаг 2. Критическая область: $\alpha = 0{,}05$, $df = 6 - 2 = 4$. Критические значения по таблице равны ±2,776.

Шаг 3. Статистика по выборке:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0{,}897\sqrt{\frac{6-2}{1-(0{,}897)^2}} = 4{,}059$$

Решение

Шаг 4. Сравниваем значение статистики с критической областью. Нулевую гипотезу отвергаем, так как значение критерия попадает в область критических значений.



Шаг 5. Делаем вывод, что существует значимая связь между признаками.

Корреляция и причинная связь

Когда проверка гипотезы показывает, что существует значимая линейная связь между переменными, исследователи должны рассмотреть возможные виды связи между переменными и выбрать ту, которая диктуется логикой данного исследования.

Пять видов связи между переменными

1. **Прямая причинно-следственная связь между переменными** (x определяет y). Наличие воды ускоряет рост растений, яд вызывает смерть, жара – таяние льда.

2. **Обратная причинно-следственная связь между переменными** (y определяет x). Исследователь может думать, что чрезмерное потребление кофе вызывает нервозность. Но, может быть, очень нервный человек хочет кофе, чтобы успокоить свои нервы?

3. **Связь между переменными вызвана третьей переменной**. Исследователь установил, что существует некая зависимость между числом утонувших людей и числом выпитых безалкогольных напитков в летнее время. Может быть, обе переменные связаны с жарой и потребностью во влаге?

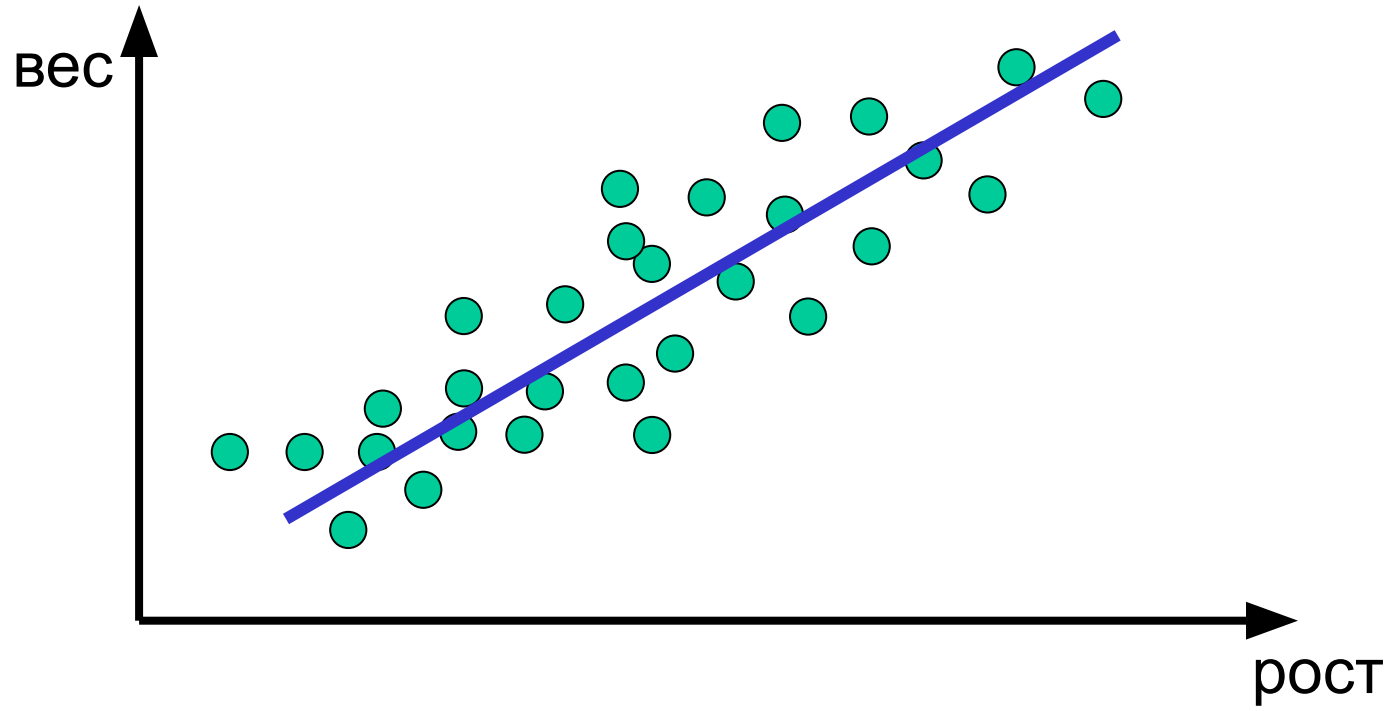
4. **Взаимосвязь между несколькими переменными**. Исследователь может обнаружить значимую связь между оценками студентов в университете и оценками в школе. Но, возможно, действуют и другие переменные: IQ, количество часов занятий, влияние родителей, мотивация, возраст, авторитет преподавателей.

5. **Зависимость случайна**. Исследователь может найти значимую зависимость между увеличением количества людей, которые занимаются спортом и увеличением количества людей, которые совершают преступления. Но здравый смысл говорит, что любая связь между этими двумя переменными должна быть случайной.

12.3. Регрессия

Исследование зависимости

На графическом изображении видно, что с увеличением роста увеличивается и вес. Зависимость имеет приближенно линейный характер. Значения переменных колеблются вокруг некоей гипотетической прямой линии, которая называется **линией регрессии**. Как её построить?



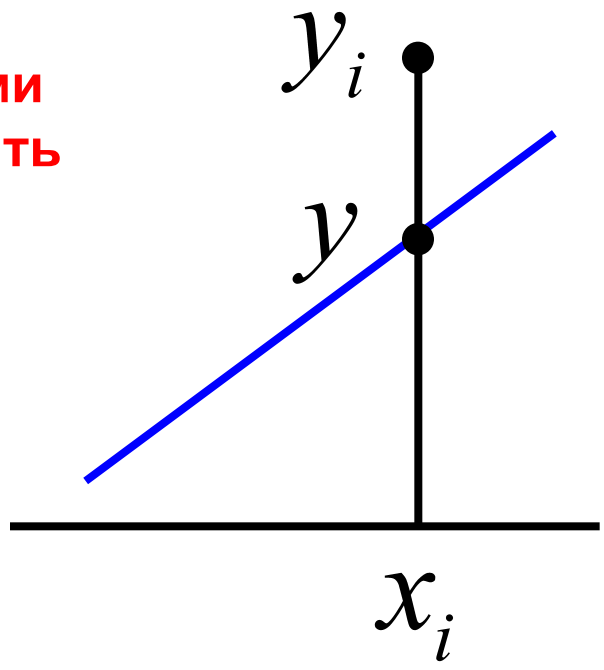
Какая прямая наилучшая?

Наши данные представляют собой пары (x, y). Тем самым, для каждого x имеется некоторое значение y. Кроме того, для каждого x существует соответствующее ему значение линейной функции y = ax + b. Сравним их.

$$x_i \Rightarrow y_i$$

Расстояние между этими значениями должно быть минимально.

$$y = ax_i + b$$



Сумма квадратов разностей минимальна…

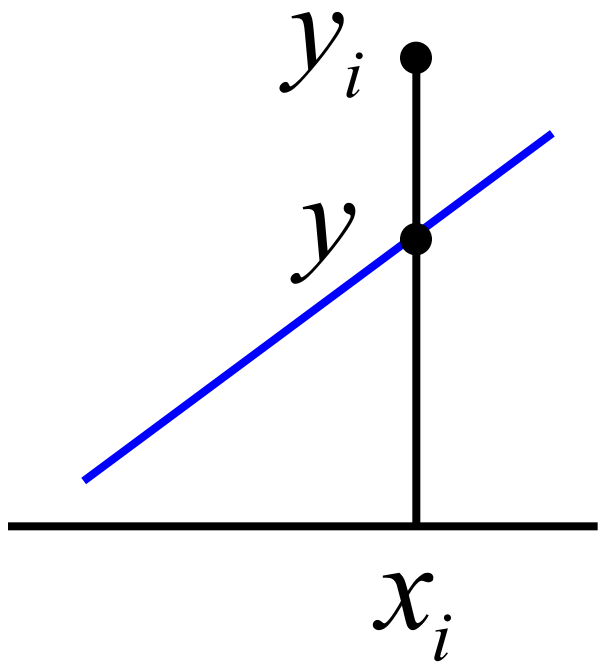
$$y_i \quad \Leftrightarrow \quad y = ax_i + b$$

Расстояние между этими значениями должно быть минимально.

$$\Updownarrow$$

$$\sum (y - y_i)^2 \to \min$$

по всем парам (x,y)



Ищем коэффициенты уравнения *y = ax + b*

В каком случае расстояние минимально?

$$\sum (y - y_i)^2 \to \min$$

Сумма зависит только от двух параметров - a и b, используем метод наименьших квадратов.

$$f(a,b) = \sum_{i=1}^{n} (y - y_i)^2 = \sum_{i=1}^{n} ((ax_i + b) - y_i)^2$$

$$\frac{\partial f(a,b)}{\partial a} = 0 \qquad \frac{\partial f(a,b)}{\partial b} = 0$$

Коэффициенты a и b

Два уравнения, которые мы получим после нахождения двух частных производных, представляют систему с двумя неизвестными. Из этой системы находятся коэффициенты, которые мы ищем:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Наклон прямой

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

Смещение прямой вдоль оси Y

Формулы для вычислений в таблице

Для табличный вычислений более удобны следующие формулы:

$$a = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$b = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

Пример вычисления

Найдем линейное уравнение регрессии для нашего примера.

Студент	Часы x	Оценка y
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

Шаг 1. Достроим таблицу

Достраиваем таблицу тремя столбцами и итоговой строкой. Проводим необходимые вычисления.

Студент	Часы x	Оценка y	xy	x^2	y^2
A	6	82	492	36	6724
B	2	63	126	4	3969
C	1	57	57	1	3249
D	5	88	440	25	7744
E	2	68	136	4	4624
F	3	75	225	9	5625
	$\Sigma x=19$	**$\Sigma y=433$**	**$\Sigma xy=1476$**	**$\Sigma x^2=79$**	**$\Sigma y^2=31935$**

Абсолютно также! То есть – можно не делать!

Шаги 2-3. Подставим в формулы, получим ответ

Подставим полученные в таблице значения в формулы для a и b*:*

$$a = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{6 \cdot 1476 - 19 \cdot 433}{6 \cdot 79 - 19^2} = 5{,}6$$

$$b = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{433 \cdot 79 - 19 \cdot 1476}{6 \cdot 79 - 19^2} = 54{,}5$$

Ответ. Получили уравнение «наилучшей прямой»:

$$y = 5{,}6\,x + 54{,}5$$

Интерпретация

1. **Увеличение времени подготовки на 1 час приводит к улучшению результата на 5,6 балла.**

2. **Чтобы улучшить результат на 10 баллов, нужно заниматься на 1,8 часа больше.**

3. Если не заниматься вообще – получишь 54,5 балла.

4. Чтобы получить 100 баллов, нужно заниматься 8,1 часов.

$$y = 5{,}6\,x + 54{,}5$$

Выходим за границы анализируемой области!

Отчет из SPSS

Отчет о расчете коэффициентов регрессии, полученный из SPSS.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	24,248	4,549		12,839	,000
	VAR00001	2,266	,174	,955	13,124	,000

a. Dependent Variable: VAR00002

Будьте осторожны с прогнозами!

Когда прогнозы распространяются за пределы исследуемых данных, интерпретировать результаты необходимо с особой осторожностью.

В 1979 году некоторые эксперты предсказывали, что в США к 2003 году запасы нефти будут исчерпаны. Этот прогноз основывался на уровне потребления нефти, характерного для того времени, и на знании объема имевшихся запасов. Однако с тех пор автомобильная промышленность выпустила много энергоемких машин. Также, существуют множество все еще неоткрытых нефтяных месторождений. Наконец, когда-нибудь наука откроет, как использовать другие виды топлива для автомобилей, что-нибудь вроде арахисового масла.

Помните, что, когда делаются прогнозы, они основываются на текущих условиях или на предположении, что существующие ныне тенденции продолжатся в будущем. Это предположение может оправдаться или не оправдаться.

12.4. Надежность прогноза

Регрессионное исследование

Уже научились:

Шаг 1. Графически изобразить пары значений (x, y).

Шаг 2. Если визуально просматривается связь, найти коэффициент корреляции.

Шаг 3. Оценить значимость коэффициента корреляции.

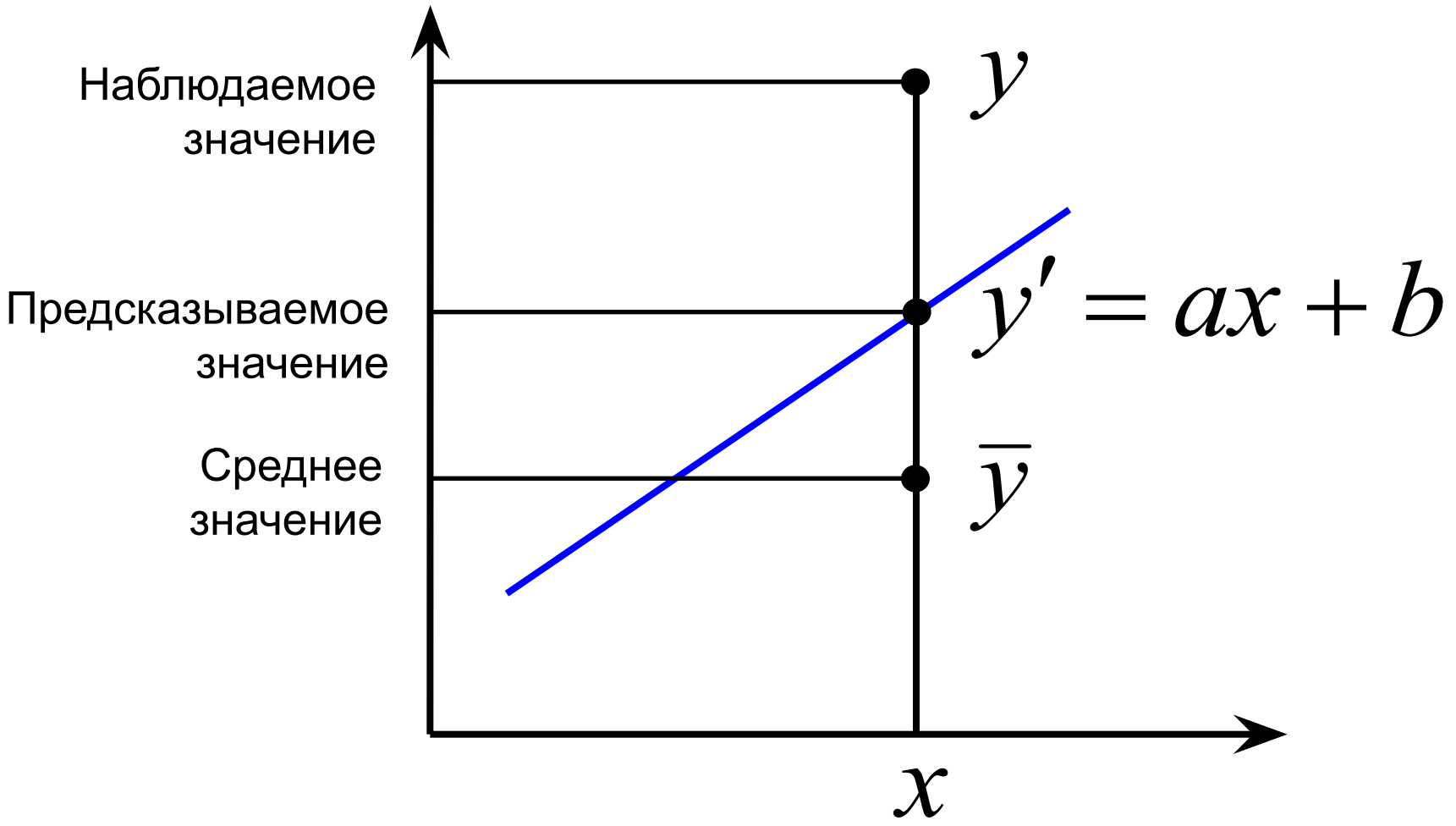
Шаг 4. Если коэффициент значим, то найти уравнение регрессии.

Шаг 5. Построить разумные прогнозы: для значения независимой переменной *x* предсказать значение зависимой переменной *y*.

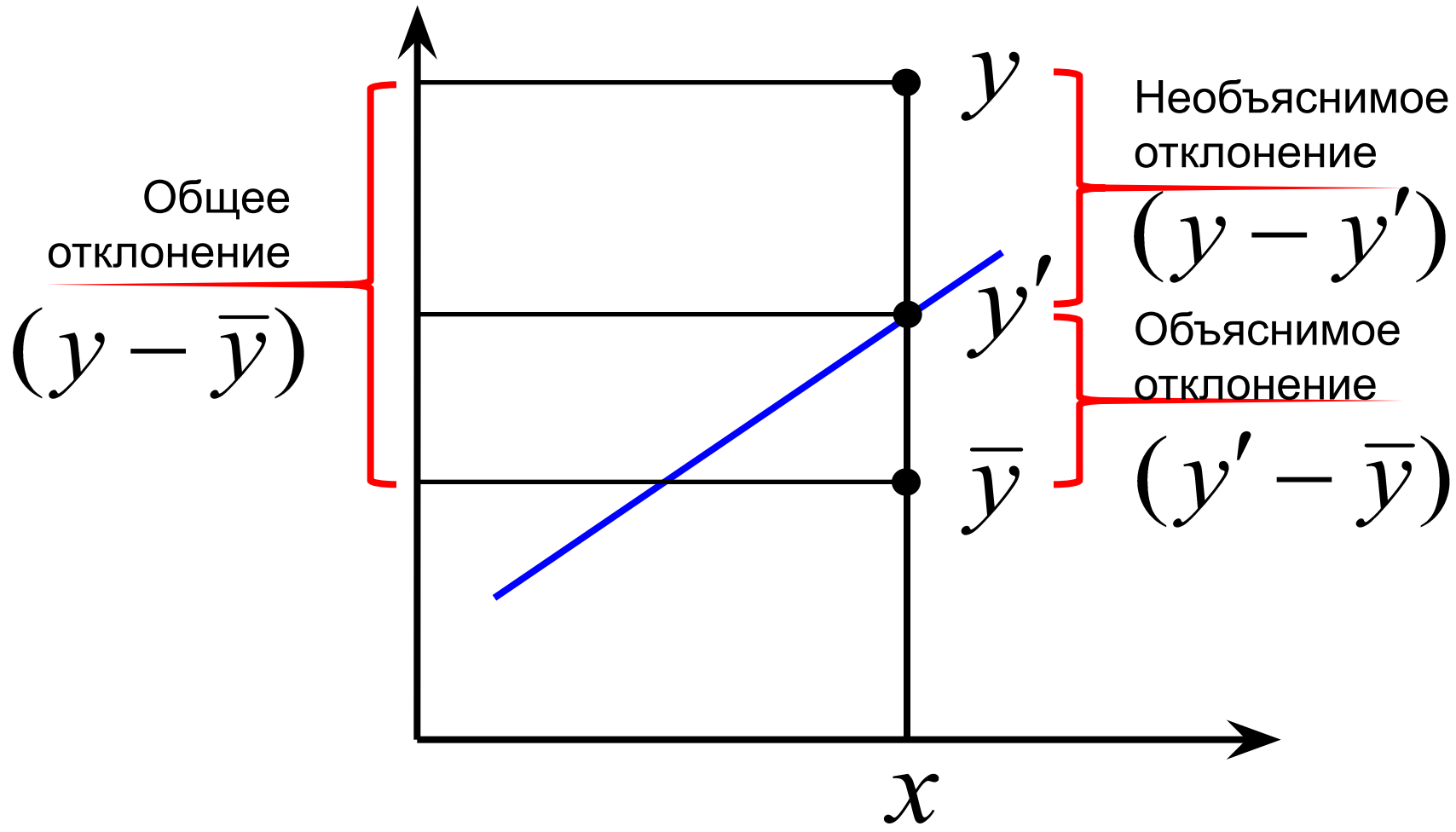
Научимся:

Шаг 6. Оценить надежность прогноза: найти коэффициент детерминации, стандартную ошибку оценки и интервал предсказания.

Наблюдаемые и предсказываемые значения



Объяснимое и необъяснимое отклонение



Вариация в регрессионной модели

Общее отклонение есть сумма объяснимой и необъяснимой вариации:

$$\sum (y - \bar{y})^2 = \sum (y' - \bar{y})^2 + \sum (y - y')^2$$

Общая вариация	Объяснимая вариация	Необъяснимая вариация

Пример

Рассчитаем общее отклонение, объяснимую и необъяснимую вариацию.

Студент	Часы x	Оценка y	y'	$(y'-\bar{y})^2$	$(y-y')^2$	$(y-\bar{y})^2$
A	6	82	87,9	248,7	35,2	96,7
B	2	63	65,7	42,2	7,1	84,0
C	1	57	60,1	145,5	9,6	230,0
D	5	88	82,4	104,1	31,7	250,7
E	2	68	65,7	42,2	5,4	17,4
F	3	75	71,2	0,9	14,2	8,0
	Σ=19	**Σ=433**		**Σ=583,5**	**Σ=103,3**	**Σ=686,8**

$$\bar{y} = 433/6 = 72{,}2$$

Коэффициент детерминации

Коэффициент детерминации вычисляется как отношение объяснимой вариации к общей вариации:

$$r^2 = \frac{\text{объяснимая вариация}}{\text{общая вариация}}$$

Коэффициент детерминации – это мера вариации зависимой переменной, которая определяется линией регрессии и независимой переменной. Коэффициент обозначается r^2.

Пример

Вычислим на основе результатов, полученных в таблице:

$$r^2 = \frac{\text{объяснимая вариация}}{\text{общая вариация}} = \frac{583{,}5}{686{,}8} = 0{,}85$$

Интерпретация коэффициента детерминации

Значение коэффициента детерминации можно получить, если возвести в квадрат коэффициент корреляции.

Если r = 0,922, то r^2 = 0,85 или 85%. Это означает, что 81% вариации зависимой переменной определяется вариацией независимой переменной.

Оставшиеся 19% – необъяснимая или случайная вариация. Это значение называется **коэффициентом недетерминации** и находится вычитанием коэффициента детерминации из единицы.

По мере того, как r приближается к нулю, значение r^2 уменьшается еще быстрее. Например, если r = 0,6, то r^2 = 0,36, то есть только 36% вариации зависимой переменной могут быть связаны с вариацией независимой переменной.

Стандартная ошибка оценки

Стандартная ошибка оценки – это стандартное отклонение наблюдаемых значений *y* от предсказываемых значений *y'*:

$$s_{\text{est}} = \sqrt{\frac{\sum (y - y')^2}{n - 2}}$$

Стандартная ошибка оценки схожа со стандартным отклонением выборки, но не использует среднее значение. Чем ближе наблюдаемые значения к предсказываемым, тем меньше стандартная ошибка оценки.

Пример

Рассчитаем стандартную ошибку оценки в нашем примере:

$$s_{\text{est}} = \sqrt{\frac{\sum (y - y')^2}{n - 2}} = \sqrt{\frac{103{,}3}{6 - 2}} \approx 5{,}08$$

Вторая формула для стандартной ошибки

Стандартную ошибку можно также вычислять по формуле:

$$s_{\text{est}} = \sqrt{\frac{\sum y^2 - b\sum y - a\sum xy}{n-2}}$$

Эта формула более пригодна для табличный вычислений.

Интервал предсказания

Когда конкретное значение *x* подставляется в уравнение регрессии, мы получаем предсказанное значение *y'*, которое является точечной оценкой для *y*. Так как это точечная оценка, трудно сказать насколько точной она является.

Возможно построить для оценки интервал предсказания. Выбрав значение α, мы получаем интервал, который с вероятностью (1 – α) содержит реальное значение *y*.

$$y' - E < y < y' + E$$

$$E = t_{\frac{\alpha}{2}} \cdot s_{\text{est}} \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{x})^2}{n\sum x^2 - \left(\sum x\right)^2}}$$

Пример

Сколько баллов получит студент, занимавшийся 4 часа?

Решение.

Шаг 1. Провели необходимые вычисления в таблице

Шаг 2. Нашли y′ = 5,6·4 + 54,5 = 76,9

Шаг 3. Нашли стандартную оценку ошибки s_{est} =5,08

Шаг 4. Нашли t-значение α=0,95 и df = 6 – 2 = 4. Получили t=2,776

Шаг 5. Нашли E:

$$E = 2{,}776 \cdot 5{,}08 \cdot \sqrt{1 + \frac{1}{6} + \frac{6 \cdot (4 - 3{,}17)^2}{6 \cdot 79 - 19^2}} = 15{,}5$$

Пример

Шаг 6. Подставили в формулу интервала:

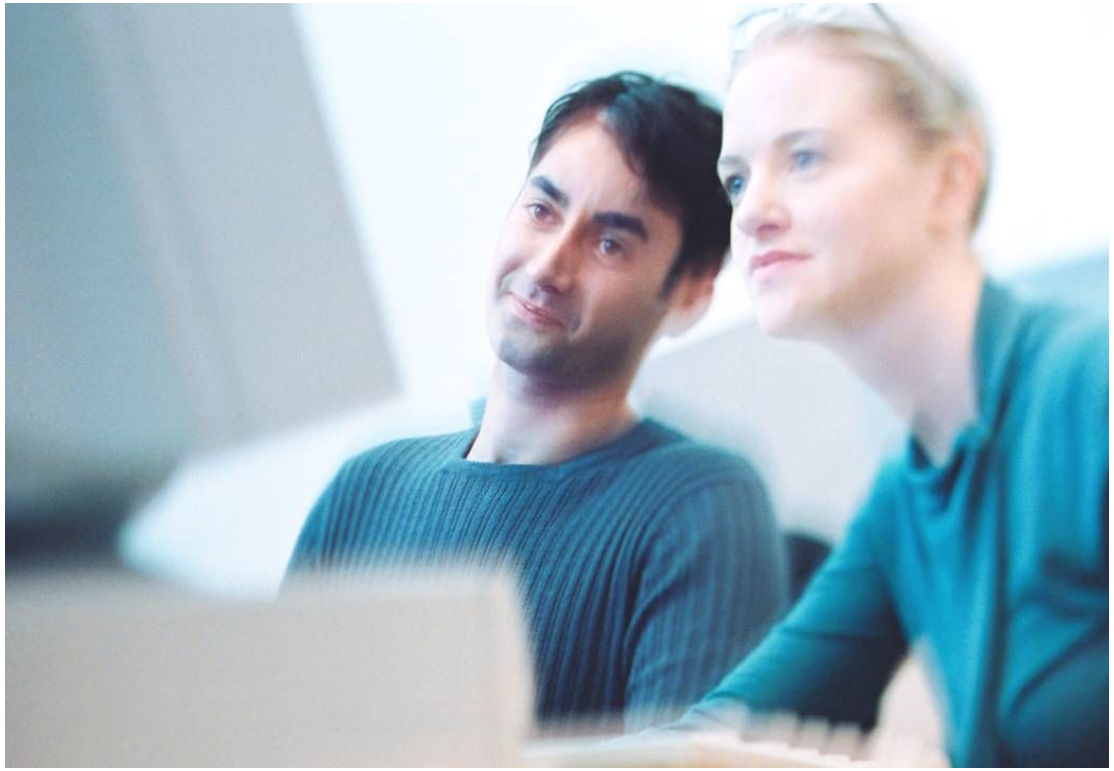
$$76{,}9 - 15{,}5 < y < 76{,}9 + 15{,}5$$

Ответ. Прогнозируемое значение баллов, которое может получить студент при 4 часах подготовки, находится с вероятностью 95% в интервале:

$$61{,}4 < y < 92{,}4$$

Задание на 5 минут

Можно ли при помощи χ^2 критерия проверить гипотезу о том, является ли распределение биномиальным? Если да, то каким образом? Если нет, то почему?



Задачи

12-1. Исследователь хочет определить, существует ли связь между возрастом человека и тем, сколько часов в день он или она смотрит телевизор.

Возраст, *x*	18	24	36	40	58
Количество часов, *y*	3,9	2,6	2	2,3	1,2

Задачи

12-2. Президент ассоциации выпускников знаменитого колледжа хочет определить, есть ли какая либо взаимосвязь между размерами вносимых бывшими учениками благотворительных пожертвований, и количеством лет, прошедших после того, как они закончили колледж.

Годы, *x*	1	5	3	10	7	6
Вклад, *y*	500	100	300	50	75	80

Задачи

12-3. Менеджер магазина хотел бы узнать существует ли какая-либо связь между возрастом работников и количеством больничных, которые они берут каждый год.

Возраст	18	26	39	48	53	58
Дни	16	12	9	5	6	2

12-4. Преподавателю необходимо узнать, насколько сильна связь между IQ студента и средним получаемым им баллом.

IQ	98	105	100	100	106	95	116	112
Средний балл	2,1	2,4	3,2	2,7	2,2	2,3	3,8	3,4

Задачи

12-5. Исследователь хочет определить, есть ли связь между тем, сколько лет уже прослужила копировальная машина, и тем, во сколько обходится ее ремонтное обслуживание в течение месяца.

Возраст	3	5	2	1	2	4	3
Стоимость обслуживания	80	100	75	60	80	93	84

Задачи

12-6. Вычислите значение *r* для следующих данных и проверьте гипотезу

$$H_0: \rho = 0$$

Нарисуйте график.

Проинтерпретируете результаты.

x	-3	-2	-1	0	1	2	3
y	9	4	1	0	1	4	9

Задачи

В задачах 12-7 по 12-10 проведите регрессионный анализ:

а) Нарисуйте график.
б) Вычислите значение коэффициента корреляции.
в) Сформулируйте нулевую и альтернативную гипотезы.
г) Проверьте их на уровне значимости $\alpha = 0{,}05$.
д) Найдите уравнение регрессии.
е) Нарисуйте линию регрессии на графике рассеивания.
ж) Сделайте выводы.

Задачи

12-7. Было проведено исследование легочных заболеваний. Полученные данные дают информацию о том, сколько лет человек курит и насколько сильно повреждены его легкие (в процентах). Сделайте прогноз относительно того, насколько будут повреждены легкие человека, который курит уже в течение 30-ти лет.

Кол-во лет, *x*	22	14	31	36	9	41	19
Повреждение легких, *y*	20	14	54	63	17	71	23

Задачи

12-8. Преподаватель статистики заинтересован в том, чтобы узнать силу связи между баллами, полученными на выпускных экзаменах студентами, проходившими обучение в первой и во второй группах по статистике. Данные в процентах в таблице.

Группа 1, *x*	87	92	68	72	95	78	83	98
Группа 2, *y*	83	88	70	74	90	74	83	99

Задачи

12-9. Преподаватель стремится понять, как число пропущенных студентом занятий влияет на его итоговый балл. Данные выборки в таблице.

Количество пропусков, *x*	10	12	2	0	8	5
Итоговый балл, *y*	70	65	96	94	75	82

Задачи

12-10. Было проведено исследование, нацеленное на то чтобы выявить, как зависит от ежемесячного дохода человека то, сколько он готов потратить на развлечения. Данные выборки (в долларах) в таблице.

Доход	800	1200	1000	900	850	907	1100
Траты на развлечения	60	200	160	135	45	90	150

Задачи

12.11. Для задачи 12.1. найдите уравнение регрессии и предскажите значение для возраста 38 лет. Найдите стандартную ошибку предсказания и найдите 90% интервал предсказания при x = 20 лет.

12.12. Для задачи 12.2. найдите уравнение регрессии и предскажите значение для 4 лет. Найдите стандартную ошибку предсказания и найдите 95% интервал предсказания при x = 4 года.

12.13. Для задачи 12.3. найдите уравнение регрессии и предскажите значение для 28 лет. Найдите стандартную ошибку предсказания и найдите 98% интервал предсказания при x = 47 лет.