

# Математические методы в биологии

## Блок 3. Математическая статистика

### Лекция 7

Козлова Ольга Сергеевна  
89276755130, [olga-sphinx@yandex.ru](mailto:olga-sphinx@yandex.ru)

# Понятие корреляции

- Взаимосвязь между количественной и качественной переменной – t-test (если качественная переменная представлена двумя градациями) или дисперсионный анализ + критерий Тьюки (если градаций больше)

ВОПРОС: А как исследовать взаимосвязь между двумя количественными переменными?

*Например, между ростом и весом, между возрастом и IQ и т.п.*

**Корреляция** – статистическая взаимосвязь двух случайных величин.

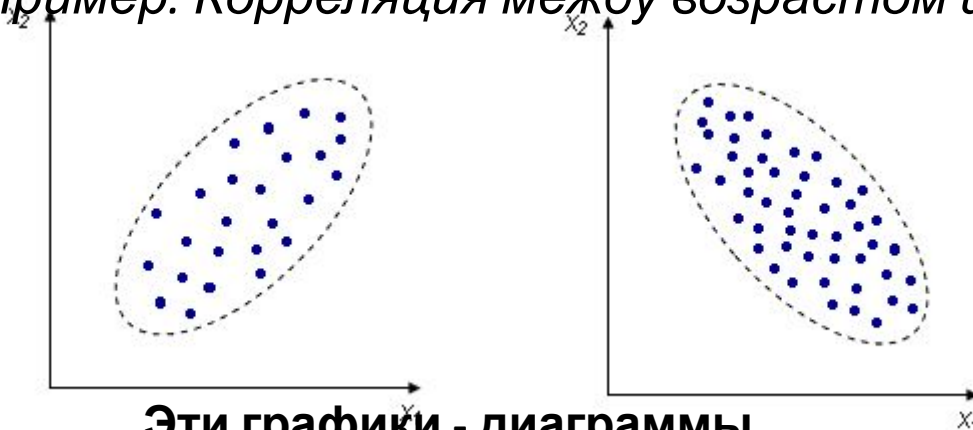
Бывает:

- Положительной

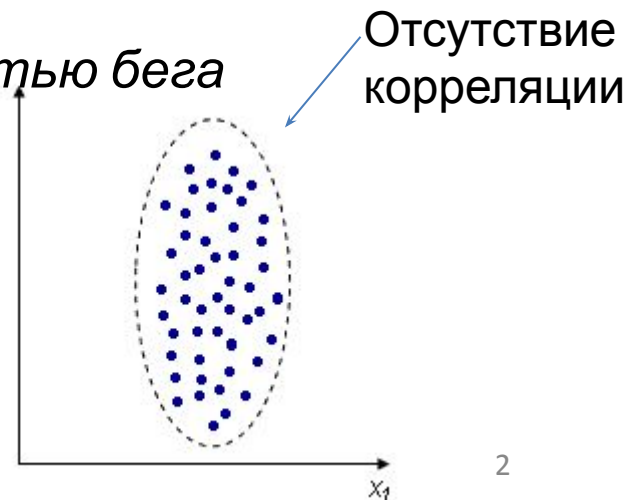
*Пример. Корреляция между ростом и весом*

- Отрицательной

*Пример. Корреляция между возрастом и скоростью бега*



Эти графики - диаграммы



# Коэффициент корреляции

- Это численный показатель, позволяющий определить:
  - направление корреляции (положительная/отрицательная)
  - её силу

Для каждого из наблюдений можно вычислить

$$(x_1 - \bar{x}_1) * (x_2 - \bar{x}_2)$$

А для всей группы – взять среднее:

$$\frac{\sum (x_1 - \bar{x}_1) * (x_2 - \bar{x}_2)}{n - 1}$$

Ковариация  
(cov)

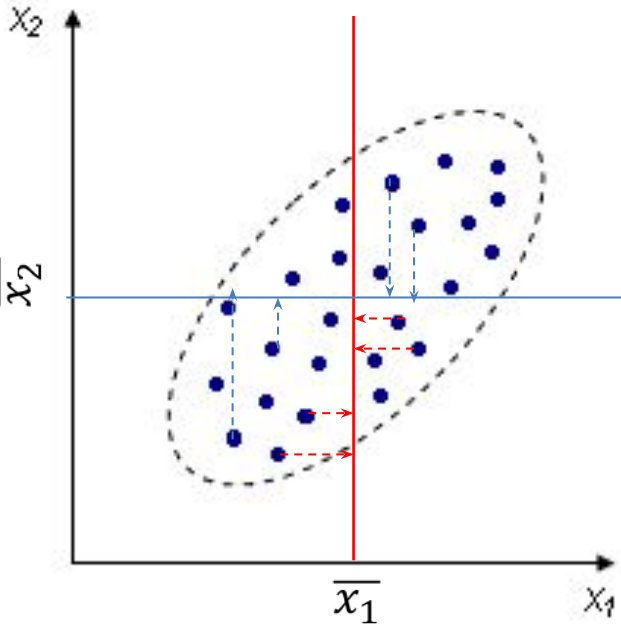
По аналогии с  
дисперсией

Величина  $cov \in (-\infty; +\infty)$  - затрудняет сравнение результатов разных экспериментов между собой, зависит от масштаба  
ВОПРОС: можем ли мы «сжать» этот интервал до [-1;+1]?

ОТВЕТ: Да, если поделим ковариацию на произведение стандартных отклонений

Коэффициент  
корреляции  
(Пирсона)

$$\longrightarrow r_{x_1x_2} = \frac{cov}{sd_{x_1} sd_{x_2}}$$



# Почему коэффициент корреляции варьирует на [-1;+1]?

- $$cov_{x_1x_2} = \frac{\sum(x_1 - \bar{x}_1) * (x_2 - \bar{x}_2)}{n - 1}$$

$$r_{x_1x_2} = \frac{cov}{sd_{x_1}sd_{x_2}} = \frac{\sum(x_1 - \bar{x}_1) * (x_2 - \bar{x}_2)}{n - 1} : \left( \frac{\sqrt{\sum(x_1 - \bar{x}_1)^2}}{\sqrt{n - 1}} * \frac{\sqrt{\sum(x_2 - \bar{x}_2)^2}}{\sqrt{n - 1}} \right)$$

$$r_{x_1x_2} = \frac{\sum(x_1 - \bar{x}_1) * (x_2 - \bar{x}_2)}{\sqrt{\sum(x_1 - \bar{x}_1)^2} * \sqrt{\sum(x_2 - \bar{x}_2)^2}}$$

Обозначим  $(x_1 - \bar{x}_1)$  как А, а  $(x_2 - \bar{x}_2)$  – как В:

$$r_{x_1x_2} = \frac{\sum AB}{\sqrt{\sum A^2} \sqrt{\sum B^2}}$$

Скалярное произведение векторов А и В  
Норма вектора  
Норма вектора<sup>В</sup>

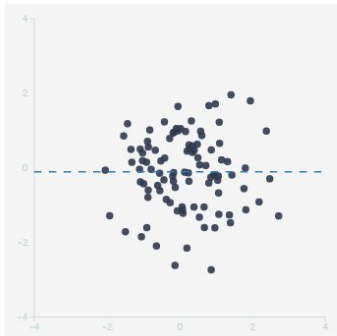
Согласно неравенству Коши-Буняковского,  $|\sum AB| \leq \sqrt{\sum A^2} \sqrt{\sum B^2}$

Отсюда  $\frac{|\sum AB|}{\sqrt{\sum A^2} \sqrt{\sum B^2}} \leq 1$ , и, значит,  $r_{x_1x_2} \in [-1; 1]$

# Коэффициент детерминации $R^2$

- Это коэффициент корреляции в квадрате
- Всегда неотрицателен и варьирует на  $[0;1]$
- $R^2$  – часть изменчивости (дисперсии) переменной, обусловленная её взаимосвязью с другой переменной

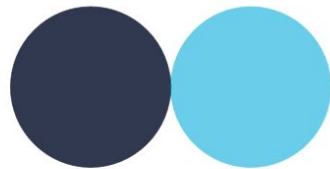
Slide me



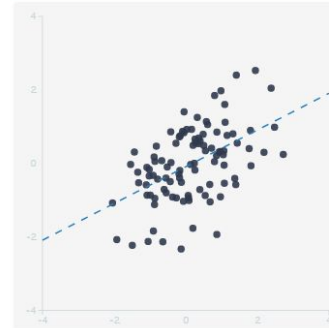
Correlation: 0

Sample size

Shared variance: 0%



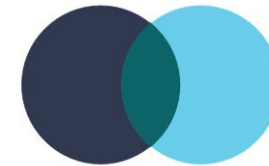
Slide me



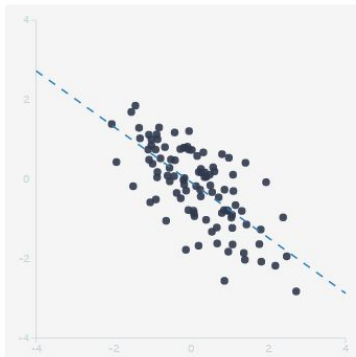
Correlation: 0.5

Sample size

Shared variance: 25%



Slide me



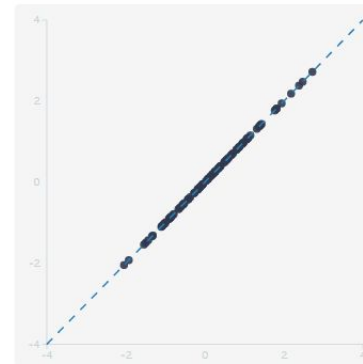
Correlation: -0.7

Sample size

Shared variance: 49%



Slide me



Correlation: 1

Sample size

Shared variance: 100%



Визуализация с сайта

# Статистическая значимость коэффициента корреляции Пирсона

- Есть две количественные переменные – X и Y. Объем выборки равен N.

$$H_0: \mu(r_{XY})=0$$

$$H_1: \mu(r_{XY})\neq 0$$

Случайная величина  $r_{XY}$  имеет t-распределение с числом степеней свободы N-2 (так как переменных две) => осталось рассчитать стандартную ошибку и можем найти уровень значимости (p-value) привычным способом.

*ВОПРОС: Всегда ли высокий коэф-т корреляции  $r_{XY}$  (напр., 0,7) будет статистически значимым?*

$$SE_{r_{XY}} = \sqrt{\frac{1 - r_{XY}^2}{N - 2}}$$

*ОТВЕТ: Нет, всё зависит от объёма выборки (числа степеней свободы)!*

$$\text{Пусть } N=50, r_{XY}=0.7. SE_{r_{XY}} = \sqrt{\frac{1-0.7^2}{48}} = 0.103. t=0.7/0.103=6.8$$

$$\text{Пусть } N=30, r_{XY}=0.7. SE_{r_{XY}} = \sqrt{\frac{1-0.7^2}{28}} = 0.135. t=0.7/0.135=5.18$$

$$\text{Пусть } N=10, r_{XY}=0.7. SE_{r_{XY}} = \sqrt{\frac{1-0.7^2}{8}} = 0.252. t=0.7/0.252=2.78$$

С  
уменьшением  
N  
уменьшается  
и t-значение

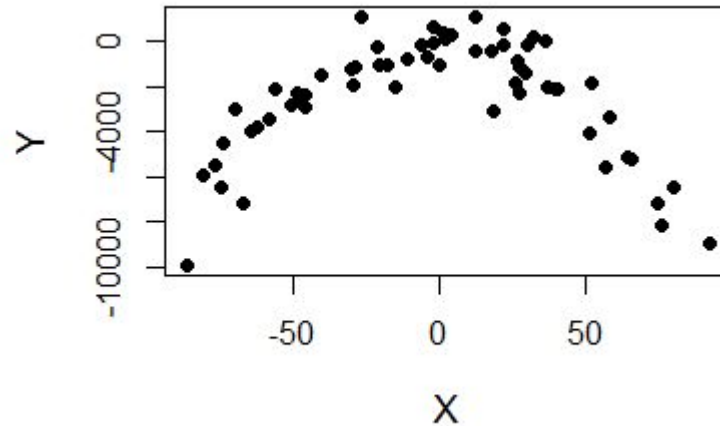
Для N=8 t-значение равно 2.4, и результат уже не статистически значим!

# Условия применения коэффициента корреляции Пирсона

- Характер взаимосвязи – прямолинейный и монотонный

Проверка. Графически – построить диаграмму рассеяния

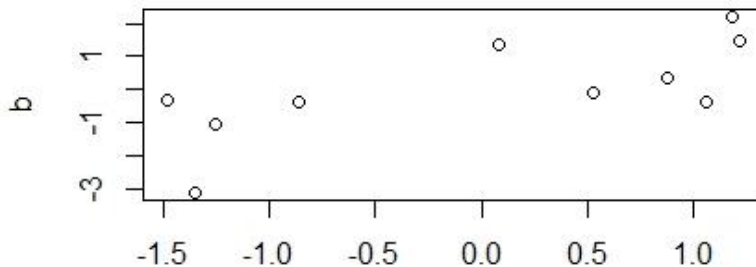
*Пример нелинейной взаимосвязи:*



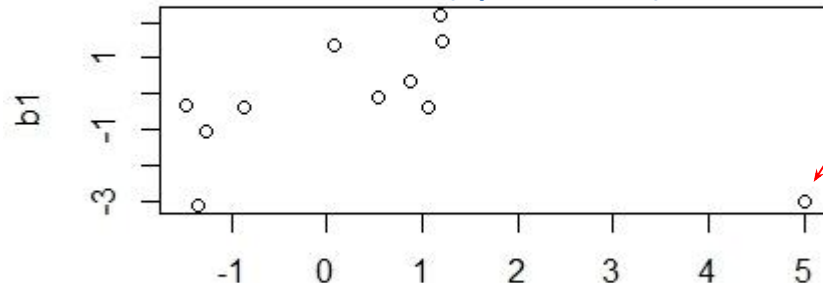
- Нормальность распределения X и Y (так как вся корреляция завязана на  $\bar{x}$ , и выбросы очень опасны)

Коэф-т кор. Спирмана – непарам. аналог

$r=0,7$  //  $r(\text{spearman})=0,67$



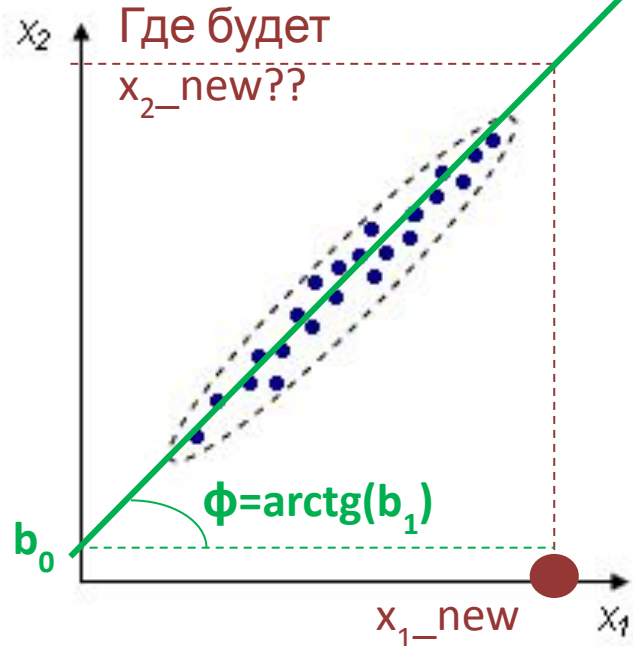
$r=-0,096$  //  $r(\text{spearman})=0,336$



выброс

# Регрессионный анализ

- Позволяет не только ответить на вопрос, есть ли взаимосвязь, но и описать, какая это взаимосвязь (построить модель взаимосвязи)
- Простейший случай – модель с одной зависимой переменной (Y) и одной независимой – предиктором (X). Обе переменных количественные.
- Неоценимое значение регрессионного анализа – возможность предсказать значение зависимой переменной по новому значению независимой, не участвовавшей в анализе.



Линия регрессии (линия тренда)

Её уравнение:  
Свободный член (intercept)  
Показывает, где прямая пересекает ось y

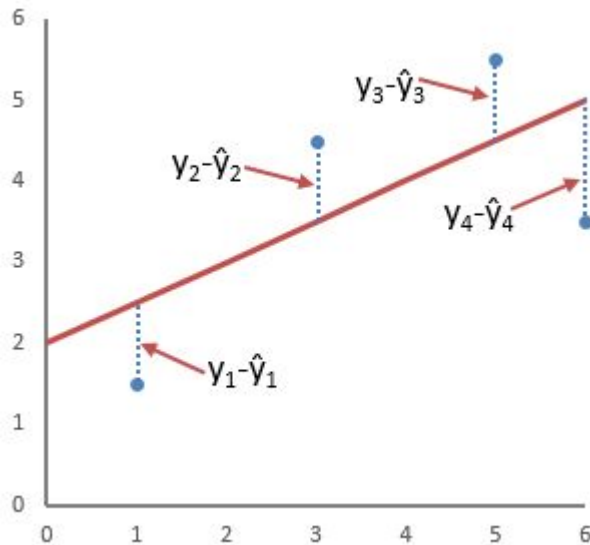
$$y = b_0 + b_1 * x$$

Коэф-т наклона (slope)  
Определяет угол наклона прямой относительно x



# Как найти оптимальную линию регрессии, или метод наименьших квадратов (МНК)

- Целевая функция – сумма квадратов остатков (разностей между фактическим и предсказанным значением зависимой переменной).
- Задача – минимизировать целевую функцию  $\sum_i (y_i - \hat{y}_i)^2$
- Те параметры линии регрессии  $b_0$  и  $b_1$ , при которых целевая функция достигает своего минимума, - оптимальны и соответствуют уравнению прямой, наилучшим образом описывающей данные.



Оптимальные параметры (без вывода):

$$b_1 = \frac{sd_y}{sd_x} * r_{xy}$$

Определяет  
знак коэф-та и  
угол наклона  
прямой

$$b_0 = \bar{y} - b_1 \bar{x}$$

$y_i$  - реальные значения переменной  $y$

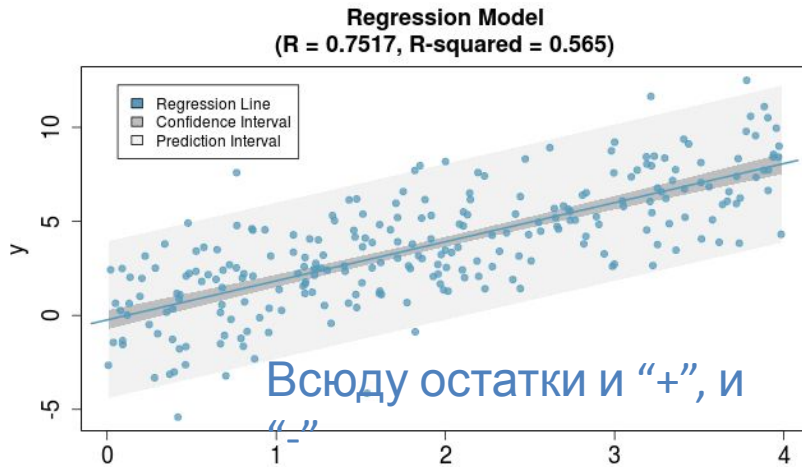
$\hat{y}_i$  - предсказанные уравнением регрессии значения

$y_i - \hat{y}_i$  - остатки

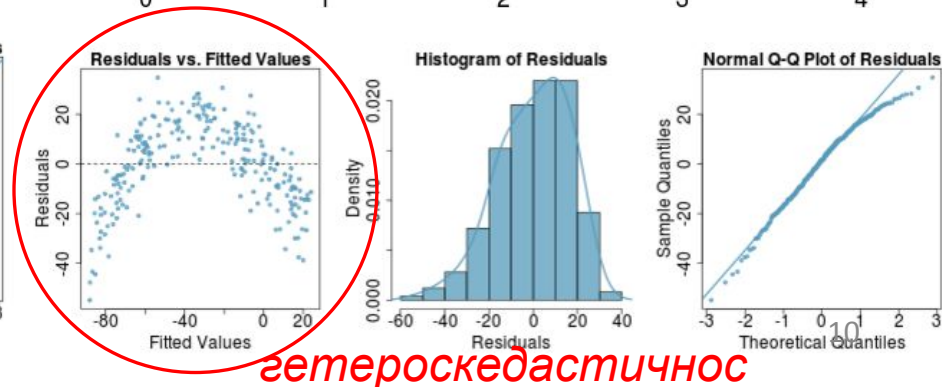
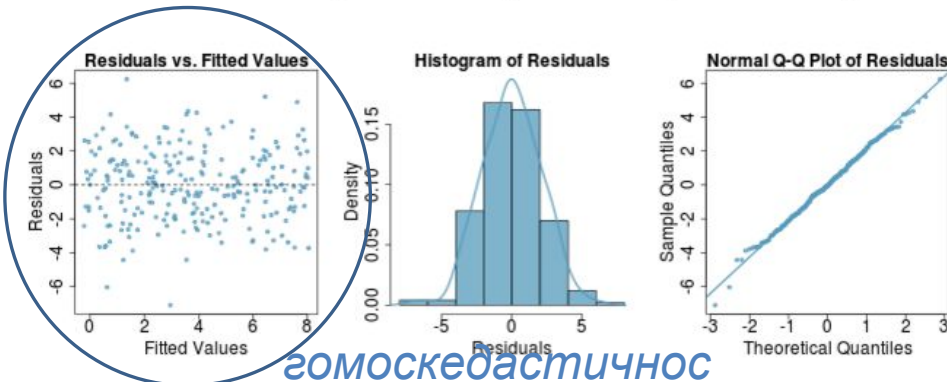
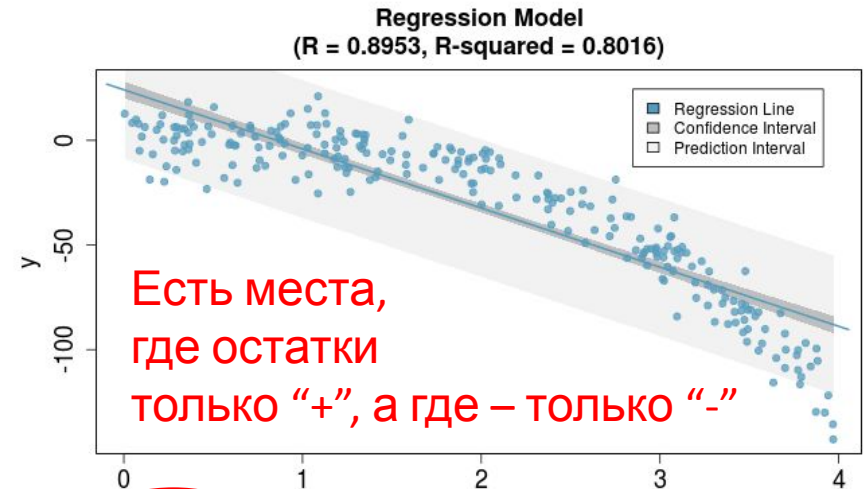
# Условия применения линейной регрессии

- Линейная взаимосвязь X и Y (проверяется диаграммой рассеяния)
- Нормальное распределение остатков  $y_i - \hat{y}_i$
- Гомоскедастичность – постоянная изменчивость остатков на всех уровнях независимой переменной

Всё хорошо:



Всё плохо:



# Пример задачи на линейную регрессию

- Исходные данные – социально-экономические показатели для штатов

	state	metro_res	white	hs_grad	poverty	female_house
1	Alabama	55.4	71.3	79.9	14.6	14.2
2	Alaska	65.6	70.8	90.6	8.3	10.8
3	Arizona	88.2	87.7	83.8	13.3	11.1
4	Arkansas	52.5	81.0	80.9	18.0	12.1
5	California	94.4	77.5	81.1	12.8	12.6
6	Colorado	84.5	90.2	88.7	9.4	9.6
7	Connecticut	87.7	85.4	87.5	7.8	12.1
8	Delaware	80.1	76.3	88.7	8.1	13.1

N=51

metro\_res - % людей, живущих в столице

white - % белокожего населения

hs\_grad - % людей с высшим образованием

poverty - % людей, живущих за чертой бедности

female\_house - % женщин-домохозяек

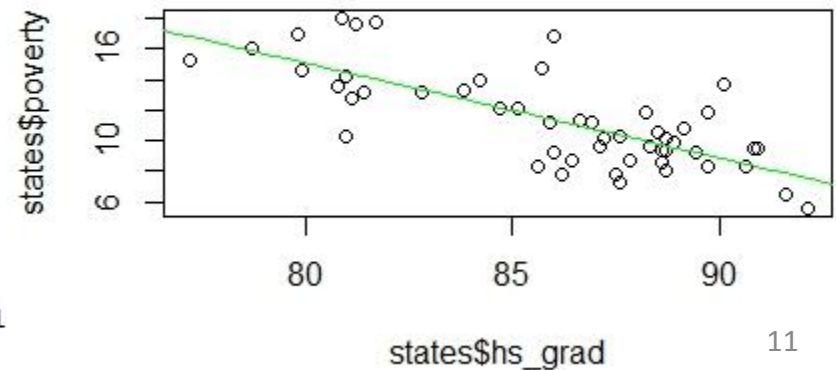
ВОПРОС: Связаны ли между собой (ко образования с уровнем бедности?

Независимая переменная – hs\_grad,  
зависимая – poverty.

Рез-ты статистически значимы

```

Coefficients:
  b0 Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.78097    6.80260    9.523 9.94e-13 ***
hs_grad      -0.62122    0.07902   -7.862 3.11e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    
```



# Ещё об интерпретации

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	64.78097	6.80260	9.523	9.94e-13	***
$b_1$ hs_grad	-0.62122	0.07902	-7.862	3.11e-10	***

signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Вероятность наблюдать t-значение, равное  $\pm 7,862$  (или выше), при условии, что верна  $H_0: \mu(b_1)=0$

Так как  $b_1 = \frac{sd_y}{sd_x} * r_{xy}$ , то это число - по сути, отражает уровень значимости найденного коэф-та корреляции (если  $sd_y \neq 0$ , то  $b_1=0$  тогда если  $r_{xy}=0$ ).

Уравнение линейной регрессии:  $\hat{y} = 64,78097 - 0,62122 * x$

%  
бедных

%  
образованных

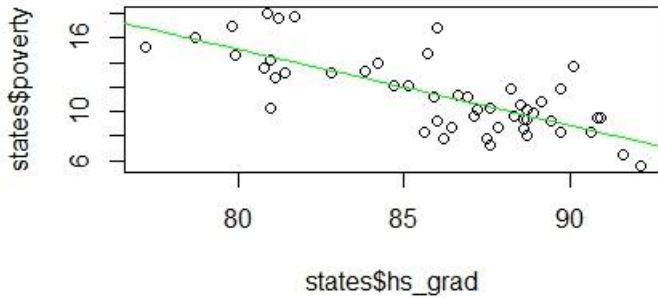
Если найдётся штат с 0% людей с высшим образованием, можно ожидать, что % людей, живущих за чертой бедности, в нём будет равен 64,78097. Далее, с каждым увеличением уровня образованности на одно деление, уровень бедности будет падать на 0,62122.

Multiple R-squared: 0.5578 – коэффициент детерминации =>  $r_{xy} = -0,747$  (корреляция отрицательна).

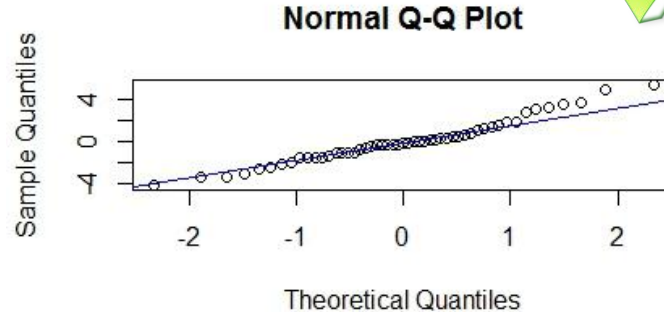
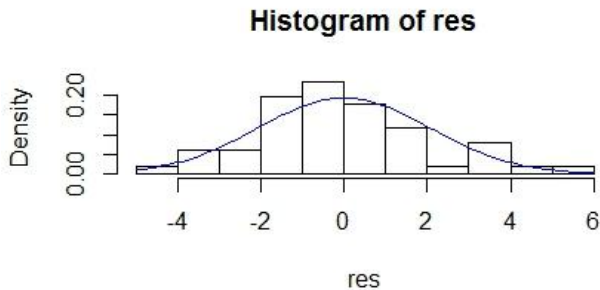
Данная линейная модель объясняет 55% изменчивости зависимой переменной.

# Наконец, проверим требования к использованию линейной регрессии

- **Линейная взаимосвязь**

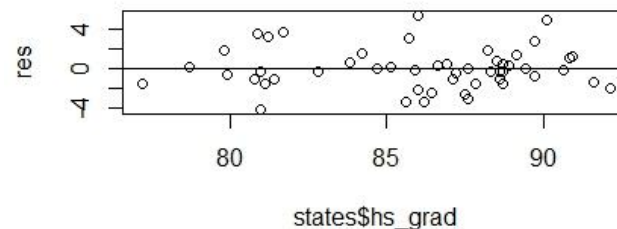


- **Нормальное распределение остатков  $y_i - \hat{y}_i$**



p-value  
(shapiro) =  
0.1831

- **Гомоскедастичность – постоянная изменчивость остатков на всех уровнях независимой переменной**



# Множественная линейная регрессия

- Несколько предикторов, одна зависимая переменная
- Уравнение регрессии выглядит так:  $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$

Зависимая  
переменная

Предиктор  
ы

- При  $n=2$  уравнение регрессии задаёт не прямую, а плоскость, а при  $n>2$  привычным образом его вообще визуализировать нельзя ☹
- Чем больше коэффициент при  $x_i$ , тем сильнее этот предиктор влияет на зависимую переменную

## ТРЕБОВАНИЯ:

Линейная взаимосвязь

Нормальное распределение остатков

Гомоскедастичность остатков

+Проверка на мультиколлинеарность (очень сильную взаимосвязь, корреляцию между какими-то из независимых переменных)

+Нормальность распределения всех переменных (желательно)

# Множественная линейная регрессия на примере

Загоним в нашу предсказательную модель для уровня бедности все оставшиеся переменные

Не оказывают влияния на зав.п.  
(коэф-ты значимо не отл. от 0)

	state	metro_res	white	hs_grad	poverty	female_house
1	Alabama	55.4	71.3	79.9	14.6	14.2
2	Alaska	65.6	70.8	90.6	8.3	10.8
3	Arizona	88.2	87.7	83.8	13.3	11.1
4	Arkansas	52.5	81.0	80.9	18.0	12.1
5	California	94.4	77.5	81.1	12.8	12.6
6	Colorado	84.5	90.2	88.7	9.4	9.6
7	Connecticut	87.7	85.4	87.5	7.8	12.1
8	Delaware	80.1	76.3	88.7	8.1	13.1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	66.47653	12.58990	5.280	3.41e-06	***
metro_res	-0.05632	0.01955	-2.881	0.006	**
white	-0.04814	0.03306	-1.456	0.152	
hs_grad	-0.55471	0.10491	-5.288	3.33e-06	***
female_house	0.05054	0.24330	0.208	0.836	

Показатели “Estimate” напротив названий переменных отражают, насколько изменится зависимая переменная с ростом данной независимой на 1 при условии, что остальные независ. пер-е зафиксированы.

При включении в модель нескольких предикторов возникает ситуация, аналогичная проблеме множественного сравнения. Поэтому имеет смысл смотреть не на сам  $R^2$ , а на его исправленную, скорректированную версию (adjusted  $R^2$ ):

Multiple R-squared: 0.6416, Adjusted R-squared: 0.6104

**Наилучшая модель – та, у которой больше всего Adjusted R-squared!**

# Проверим мультиколлинеарность

Корреляции независимых переменных между собой:

	white	hs_grad	metro_res	female_house
white		0,24	-0,34	-0,75
hs_grad			0,018	-0,62
metro_res				0,3

Переменная `female_house` сильно коррелирует с переменными `white` и `hs_grad`. Давайте удалим её из нашей модели!

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 68.72202 6.38893 10.756 2.89e-14 ***
states$metro_res -0.05553 0.01898 -2.926 0.00528 **
states$hs_grad -0.56972 0.07527 -7.569 1.13e-09 ***
states$white -0.05333 0.02148 -2.483 0.01665 *
```

Стат.значимы все 3  
независ.пер-е

```
---
signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 1.915 on 47 degrees of freedom

Multiple R-squared: 0.6412, Adjusted R-squared: 0.6183 (немного больше, чем до этого)



# Введение в логистическую регрессию

- Интересный подвид регрессии, в которой зависимая переменная – номинативная (качественная) с двумя градациями, а независимые – количественные или качественные

*Пример. Как связаны между собой средний балл по предметам в школе (количественная переменная) с тем, поступил студент в университет или нет (номинативная с двумя градациями: «0» – «не поступил», «1» – «поступил»)?*

ВОПРОС: как примирить между собой левую и правую часть уравнения регрессии  $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$ , ведь теперь у нас слева – номинативная, а справа – количественная переменная, варьирующая на  $(-\infty; +\infty)$ ?

ОТВЕТ: подменим номинативную переменную вероятностью положительного исхода (вероятностью сдачи экзамена, например)!

ВОПРОС: а как теперь «сжать» область значений в правой части, чтобы  $(-\infty; +\infty)$  превратить в  $[0;1]$  (ведь так варьирует вероятность)?

ОТВЕТ: Никак!

Но это не повод расстраиваться, ведь наша регрессия не зря называется «логистической»...

# От вероятности к логарифму шанса

- Шанс (odds) – отношение вероятности успеха к вероятности неудачи

$$odds = \frac{p}{1-p} = \frac{\text{число успехов}}{\text{число неудач}}$$

Заметим, что шанс варьирует уже на  $[0; +\infty]$ .

- А теперь рассчитаем натуральный логарифм шанса!

Таким образом, теперь и в левой, и в правой части уравнения  $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$  - действительные числа, варьирующие на  $(-\infty; +\infty)$ .

Более того, если  $\ln(odds) < 0$  (т.е.  $odds < 1$ ), то вероятность неудачи выше вероятности успеха, а если  $\ln(odds) > 0$ , то вероятность успеха выше вероятности неудачи.

**ЗАДАЧА.** Дано распределение слушателей курса по биоинформатике по полу и основной специальности. Рассчитать логарифм шанса, что случайный человек из этой выборки – биолог.

	Юнош и	Девушки	Всег о
Биологи	15	9	24
Информатик и	11	6	17
Всего	26	15	41

**РЕШЕНИЕ.** Число успехов (человек – биолог) равно 24, число неудач (человек – информатик) равно 17. Отсюда

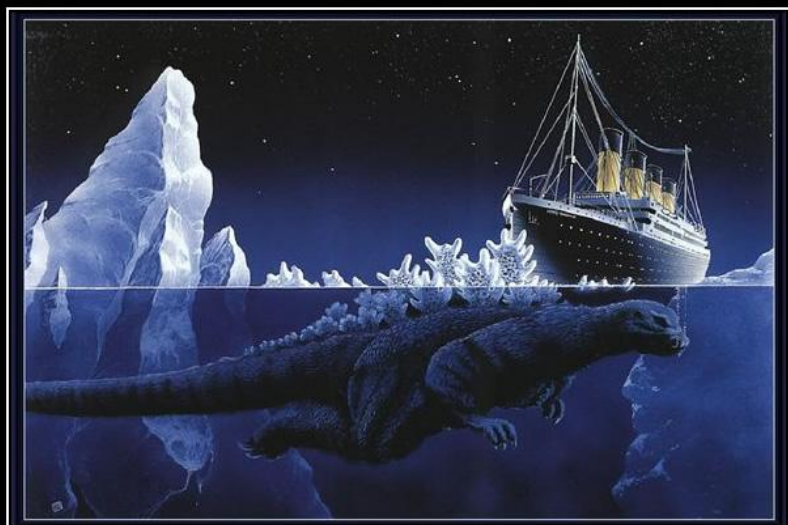
$$odds = \frac{24}{17}$$
$$\ln(odds) = 0,3448$$

# Как подбирать коэффициенты логистической регрессии?

- Будем двигаться последовательно, и начнём с модели вовсе без предикторов (intercept-only model).

В качестве тренировочного примера возьмём данные про пассажиров «Титаника» (714 наблюдений). Номинативные переменные:

- Выжил/нет (это будет зависимая переменная)
- Пол (мужчина/женщина)
- Класс каюты (1й класс/2й класс/3й класс)



Intercept в логистической регрессии – логарифм шанса успеха без учёта предикторов.

Известно, что из 714 человек выжило 290, а погибло – 424. Значит,

$$\ln\left(\frac{290}{424}\right) = -0,38$$

Это и будет первый (и единственный!) коэффициент  $b_0$ .

# Модель с одним номинативным предиктором

- Теперь будем учитывать ещё и пол пассажира.

Распределение пассажиров по полу и исходу пребывания на Титанике (таблица сопряжённости):

	Мужчин	Женщин
Выжил	93	197
Нет	360	64

Рассчитаем шанс выжить для мужчин и женщин по отдельности:

$$\text{odds}(\text{male}) = 93/360 = 0,26$$

$$\text{odds}(\text{female}) = 197/64 = 3,08$$

Их логарифмы:  $\ln(\text{odds}(\text{male})) = -1,35$

$$\ln(\text{odds}(\text{female})) = 1,12$$

*Какая градация будет базовым уровнем – выбирается просто по алфавиту!*

Отношение шансов выжить для мужчин и женщин =  $0,26/3,08 = 0,08$

Его логарифм:

$$\ln(\text{odds}(\text{male})/\text{odds}(\text{female})) = \ln(\text{odds}(\text{male})) - \ln(\text{odds}(\text{female})) = -2,47$$

Уравнение регрессии примет вид:

$$\ln(\text{odds}(\text{survive})) = 1,12 - 2,47 * \text{Sex\_male}$$

Логарифм шанса выжить, если пассажир - женщина

«Штраф» (цена перехода), если пассажир мужчина, – логарифм отношения шансов выжить для мужчин и базового уровня фактора (женщин)

Переменная, принимающая значение 0, если пассажир – женщина, и 1 – если мужчина

# Если независимая переменная - количественная

ЗАДАЧА. Исследовать, как влияет средний балл абитуриента в школе на вероятность его поступления в ВУЗ.

Исходные данные – 400 наблюдений вида

завис.пер-я (1 – поступил, 0 – нет)

независ.колич.пер-я (сред.балл в школе,  $gpa \in [2,26;4]$ )

	admit	gre	gpa	rank
1	0	380	3.61	3
2	1	660	3.67	3
3	1	800	4.00	1
4	1	640	3.19	4
5	0	520	2.93	4
6	1	760	3.00	2
7	1	560	2.98	1
8	0	400	3.08	2
9	1	540	3.39	3
10	0	700	3.92	2
11	0	800	4.00	4

Showing 1 to 11 of 400 entries

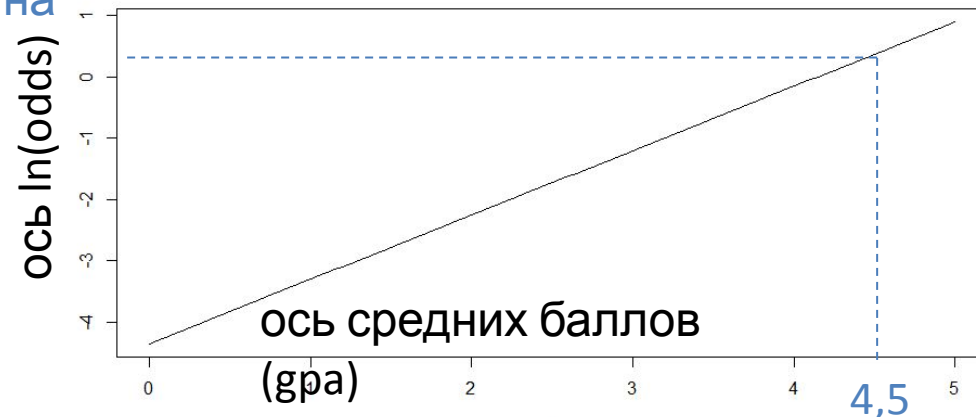
Коэффициенты уравнения регрессии:

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-4.3576	1.0353	-4.209	2.57e-05 ***
gpa	1.0511	0.2989	3.517	0.000437 ***

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Логарифмы шансов. 1,0511 – насколько увеличится логарифм шанса поступления при увеличении  $gpa$  на



Если ср.балл  $\approx 4.5$ , то  $\ln(p/(1-p)) \approx 0,3$

Отсюда  $p/(1-p) \approx \exp(0,3) \approx 1,35$

Отсюда  $p \approx 0,57$

Вероятность поступить со

# Схема анализа количественных данных

