Введение в анализ ассоциации количественных признаков

Юрий Аульченко

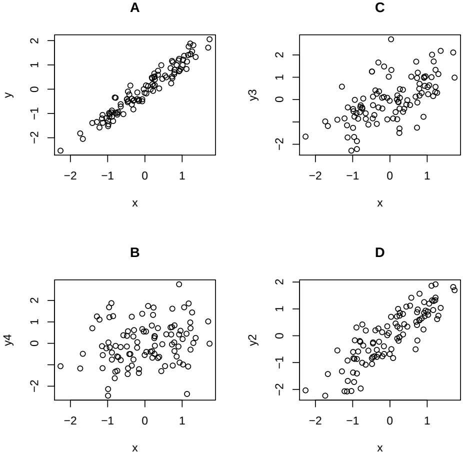
yurii [dot] aulchenko [at] gmail [dot] com

16 октября 2011 г.

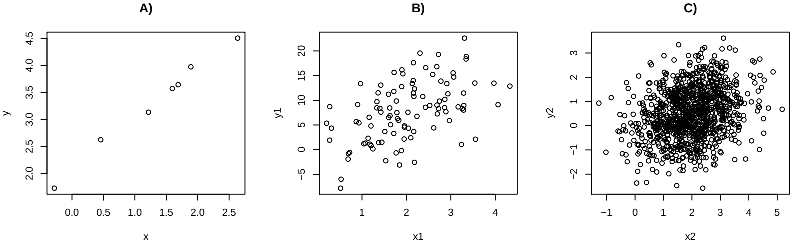
Содержание

Содержание

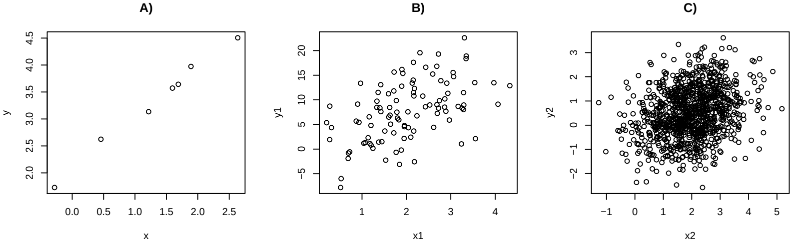
Где ассоциация сильнее?



Где ассоциация сильнее?

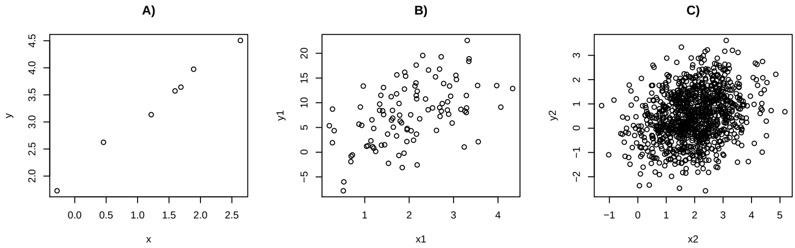


Где ассоциация сильнее?



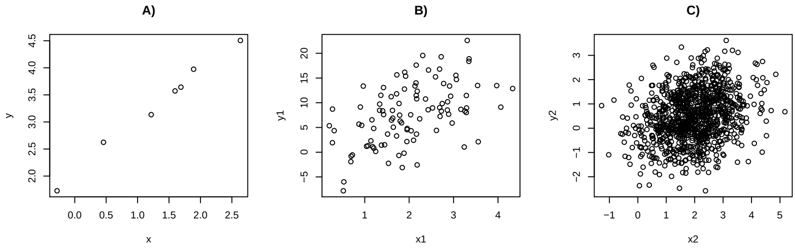
- Кажется, $A > B > C$ (?)

Где ассоциация сильнее?



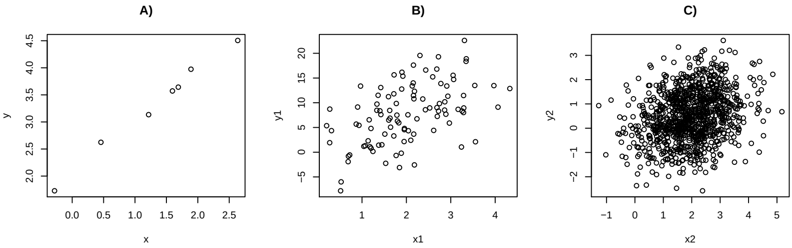
- Кажется, $A > B > C$ (?)
- Для того, что бы дать обоснованный ответ, необходимо ввести меру силы ассоциации между переменными

Где ассоциация сильнее?



- Кажется, $A > B > C$ (?)
- Для того, что бы дать обоснованный ответ, необходимо ввести меру силы ассоциации между переменными
- Как насчет того, что бы использовать коэффициент регрессии y на x?

Где ассоциация сильнее?



- Кажется, $A > B > C$ (?)
- Для того, что бы дать обоснованный ответ, необходимо ввести меру силы ассоциации между переменными
- Как насчет того, что бы использовать коэффициент регрессии y на x?
- Все согласны с тем, что коэффициент регрессии $A > B > C$?

Содержание

Модель линейной регрессии

- Обозначим зависимую переменную как y и независимую как x. Пусть y_i и x_i обозначают значение этих переменных для i-го образца
- Предположим, что значение независимой переменной задается уравнением

$$y_i = \mu + \beta \cdot x_i + \epsilon_i,$$

где μ – некоторая константа (отступ), β – коэффициент регрессии y на x, а ϵ – остаточный случайный шум

Модель линейной регрессии

- Оценки параметров μ и β подбираются таким образом, чтобы предсказание на основе модели

$$\hat{y}_i = \hat{\mu} + \hat{\beta} x_i$$

максимально соответствовало наблюдаемым данным

- В случае одной независимой переменной

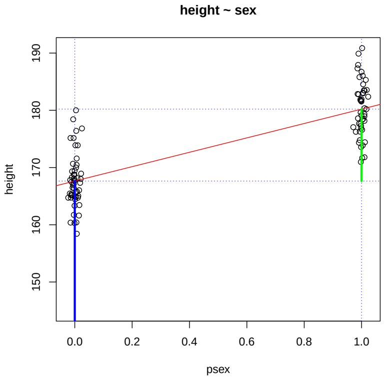
$$\hat{\beta} = \frac{Cov(x,y)}{Var(x)} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2},$$

где $\bar{x}$ и $\bar{y}$ – средние значения x и y, соответственно

Интерпретация коэффициента регрессии

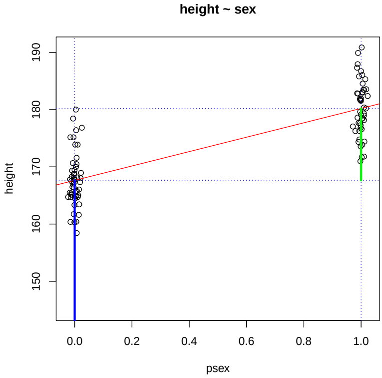
- Как отступ, так и коэффициент регрессии имеют ясную физическую интерпретацию
- Отступ μ является ожидаемым значением y когда независимая переменная x равна нулю
- Коэффициент регрессии β показывает, насколько изменяется значение y если значение x увеличивается на единицу

Пример оценивания коэффициента регрессии



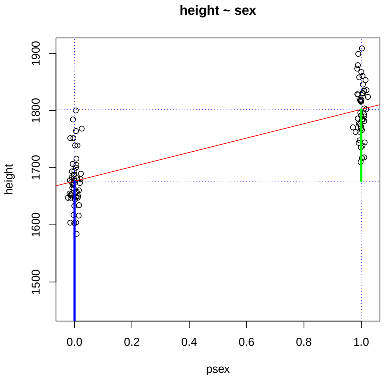
- Рассмотрим регрессионную модель $y \sim \mu + \beta \cdot x$, где y – рост в см., а x – пол ($0 =$ женский, $1 =$ мужской)
- Для выборки, состоящей из 48 мужчин и 52 женщин, были получены следуюшие оценки: $\{\hat{\mu} = 167.6, \hat{\beta} = 12.6\}$ (см. рис.)

Пример интерпретации коэффициента регрессии



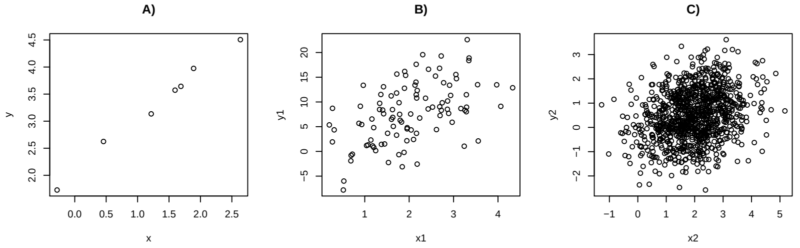
- $\hat{\mu} = 167.6$: Если $x = 0$, ожидаемое значение y равно 167.6. Другими словами, ожидаемый роста женщин равен 167.6.
- $\hat{\beta} = 12.6$: если x изменяется на 1, ожидается, что y изменится на 12.6. Другими словами, ожидаемая разница роста мужчин и женщин равна 12.6; или рост мужчин равен $\hat{\mu} + \hat{\beta} = 180.2$

Коэффициент регрессии задан на определенной шкале



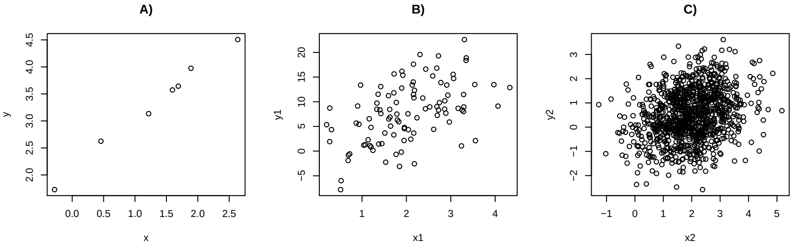
- Пусть рост измерен в милиметрах
- Тогда для того же набора данных будут получены оценки $\{\hat{\mu} = 1676, \hat{\beta} = 126\}$
- Измерение роста в мм эквивалентно умножению оценок параметров модели на 10
- Данные те же самые – используется разная шкала

Где ассоциация сильнее?



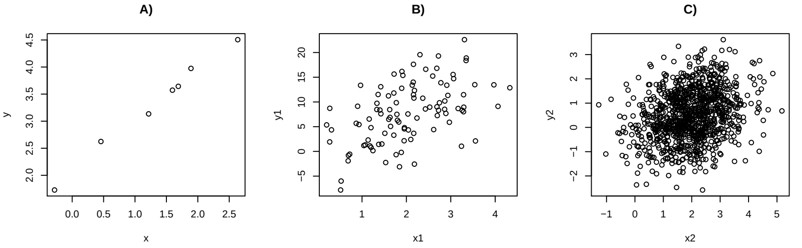
- Сила ассоциации $A > B > C$ (?)

Где ассоциация сильнее?



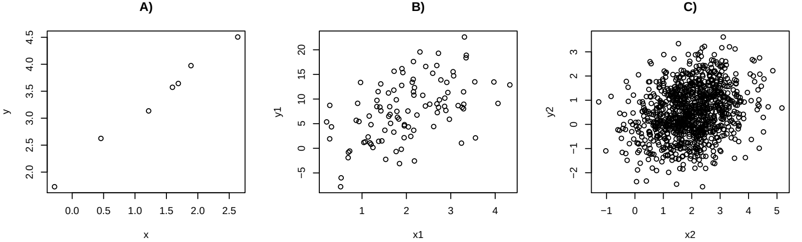
- Сила ассоциации $A > B > C$ (?)
- Коэффициент регрессии $A > B > C$?

Где ассоциация сильнее?



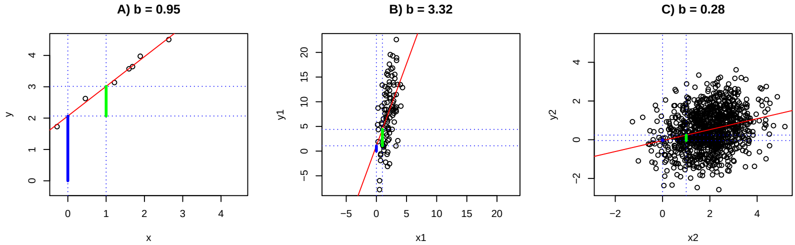
- Сила ассоциации $A > B > C$ (?)
- Коэффициент регрессии $A > B > C$?
- Коэффициент регрессии задан на определенной шкале; его можно произвольно варьировать, изменяя шкалу

Где ассоциация сильнее?



- Сила ассоциации $A > B > C$ (?)
- Коэффициент регрессии $A > B > C$?
- Коэффициент регрессии задан на определенной шкале; его можно произвольно варьировать, изменяя шкалу
- $\hat{\beta}_A = 0.95$, $\hat{\beta}_B = 3.32$ and $\hat{\beta}_C = 0.28$, так что $B > A > C$

Где ассоциация сильнее?



- Сила ассоциации $A > B > C$ (?)
- Коэффициент регрессии $A > B > C$?
- Коэффициент регрессии задан на определенной шкале; его можно произвольно варьировать, изменяя шкалу
- $\hat{\beta}_A = 0.95$, $\hat{\beta}_B = 3.32$ and $\hat{\beta}_C = 0.28$, так что $B > A > C$

Насколько хорошо x "связано" с y?

- Необходима мера ассоциации, независимая от шкалы

Насколько хорошо x "связано" с y?

- Необходима мера ассоциации, независимая от шкалы
- Мы видели, что коэффициент регрессии меняется в зависимости от шкалы – чем больше разброс независимой переменной, тем больше коэффициент

Насколько хорошо x ”связано” с y?

- Необходима мера ассоциации, независимая от шкалы
- Мы видели, что коэффициент регрессии меняется в зависимости от шкалы – чем больше разброс независимой переменной, тем больше коэффициент
- Ожидание того, на сколько изменится y при изменении x на еденицу

$$\hat{\beta}_{y \sim x} = \frac{Cov(x, y)}{Var(x)}$$

Насколько хорошо x "связано" с y?

- Необходима мера ассоциации, независимая от шкалы
- Мы видели, что коэффициент регрессии меняется в зависимости от шкалы – чем больше разброс независимой переменной, тем больше коэффициент
- Ожидание того, на сколько изменится y при изменении x на еденицу
$$\hat{\beta}_{y\sim x} = \frac{Cov(x,y)}{Var(x)}$$
- Точно так же, ожидание того, на сколько изменится x (!) при изменении y на единицу определено выражением
$$\hat{\beta}_{x\sim y} = \frac{Cov(x,y)}{Var(y)}$$

Коэффициент корреляции Пирсона

- Независимость от шкалы может быть достигнута с помощью шкалирования коэффициента регрессии $\beta_{y \sim x}$ на дисперсию y, что дает коэффициент корреляции Пирсона:

$$\rho_{xy} = \frac{Cov(x, y)}{\sqrt{Var(x) \cdot Var(y)}} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2}}$$

Коэффициент корреляции Пирсона

- Независимость от шкалы может быть достигнута с помощью шкалирования коэффициента регрессии $\beta_{y\sim x}$ на дисперсию y, что дает коэффициент корреляции Пирсона:

$$\rho_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x)\cdot Var(y)}} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\cdot\Sigma(y_i-\bar{y})^2}}$$

- Когда
  - $\rho_{xy}=1$, линейная зависимость между x и y, притом y увеличивается с увеличением x
  - $\rho_{xy}=-1$ линейная зависимость, притом y уменьшается с увеличением x
  - $\rho_{xy}=0$, нет (по крайней мере, линейной) зависимости

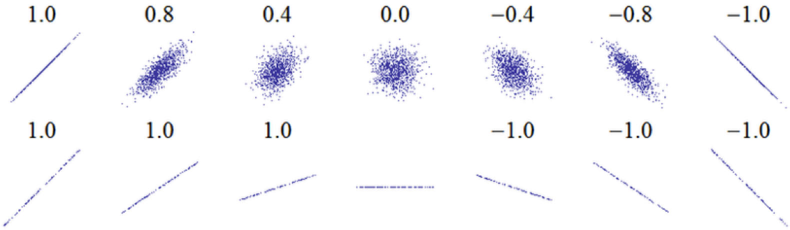
Коэффициент корреляции Пирсона

- Независимость от шкалы может быть достигнута с помощью шкалирования коэффициента регрессии $\beta_{y\sim x}$ на дисперсию y, что дает коэффициент корреляции Пирсона:

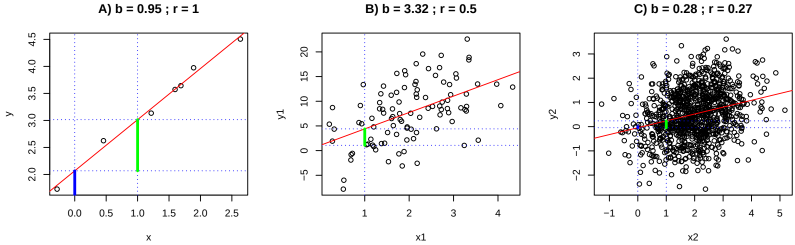
$$\rho_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x)\cdot Var(y)}} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\cdot\Sigma(y_i-\bar{y})^2}}$$

- Когда
  - $\rho_{xy} = 1$, линейная зависимость между x и y, притом y увеличивается с увеличением x
  - $\rho_{xy} = -1$ линейная зависимость, притом y уменьшается с увеличением x
  - $\rho_{xy} = 0$, нет (по крайней мере, линейной) зависимости
- Коэффициент детерминации, $\rho^2_{xy} = \beta_{y\sim x}\cdot\beta_{x\sim y}$, равен доле дисперсии y, объясненной x (и наоборот)

Примеры корреляций

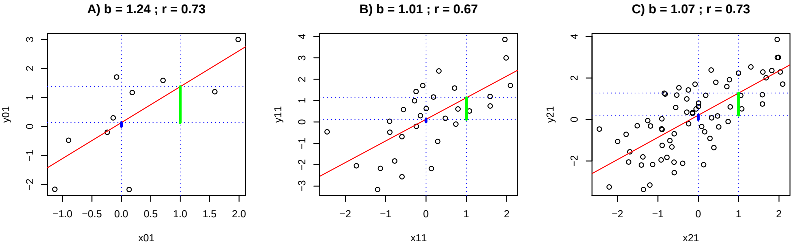


Корреляции



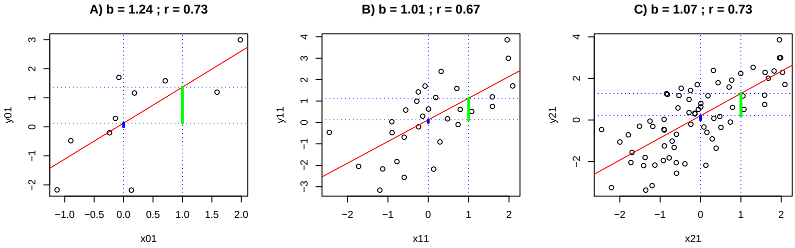
- Сила ассоциации $A > B > C$ (?)
- Регрессия: $\hat{\beta}_A = 0.95$, $\hat{\beta}_B = 3.32$ and $\hat{\beta}_C = 0.28$ $(B > A > C)$
- Корреляция: $\hat{\rho}_A = 1$, $\hat{\rho}_B = 0.5$ and $\hat{\rho}_C = 0.27$ $(A > B > C!)$

Ещё один аспект ассоциации



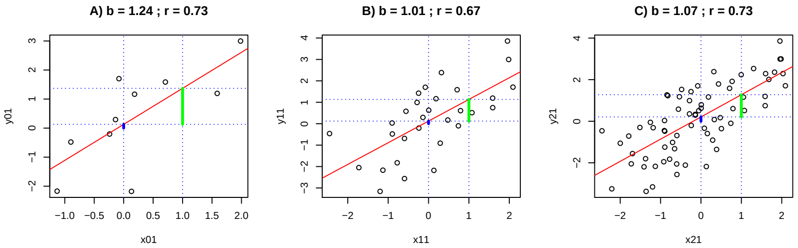
- Для A, B и C коэффициенты регрессии (и корреляции) сходны
- Значит ли это, что сила ассоциации тоже одинакова?

Ещё один аспект ассоциации



- Для A, B и C коэффициенты регрессии (и корреляции) сходны
- Значит ли это, что сила ассоциации тоже одинакова?
- Что изменяется между A, B, и C?

Корреляции



- На графике A представлено 10 точек, а на графиках B и C – 30 и 70 точек, соответственно
- В то время как величина ассоциации более-менее одинакова, степень уверенности в том, что ассоциация действительно присутствует – разная

Статистическая значимость

- Ассоциацию можно охарактеризовать задав вопрос ”какова вероятность получения такой (или даже большей) ассоциации случайным образом?”

Статистическая значимость

- Ассоциацию можно охарактеризовать задав вопрос "какова вероятность получения такой (или даже большей) ассоциации случайным образом?"
- Эта вероятность называется p-value. Чем ниже p-value, тем меньше вероятность того, что наблюденная ассоциация случайна, и тем выше статистическая значимость полученной ассоциации

Скор-тест

- Для получения p-value мы можем использовать скор-тест

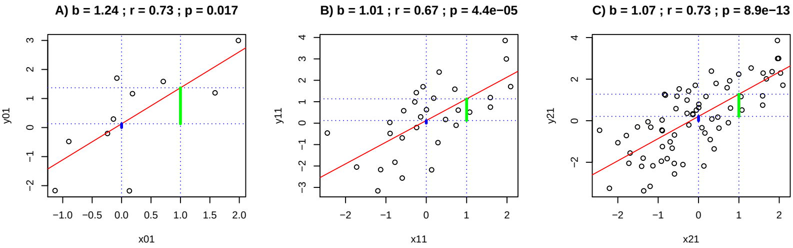
$$T^2 = \hat{\rho}_{xy}^2 \cdot n,$$

где $\hat{\rho}_{xy}^2$ – коэффициент детерминации, а n – объем выборки

Скор-тест

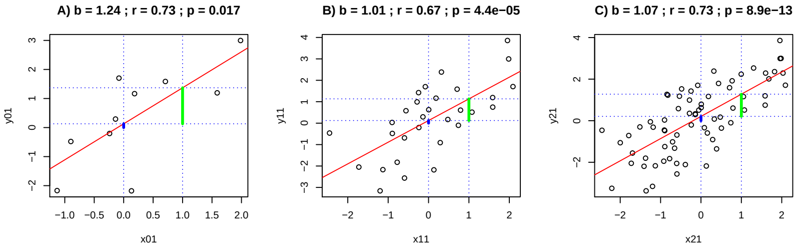
- Для получения p-value мы можем использовать скор-тест
$$T^2 = \hat{\rho}_{xy}^2 \cdot n,$$
где $\hat{\rho}_{xy}^2$ – коэффициент детерминации, а n – объем выборки
- При нулевой гипотезе об отсутствии ассоциации этот тест распределен как χ_1^2, так, что $T^2 > 3.84$ соответствует $p < 0.05$ и т.д.

Статистическая значимость



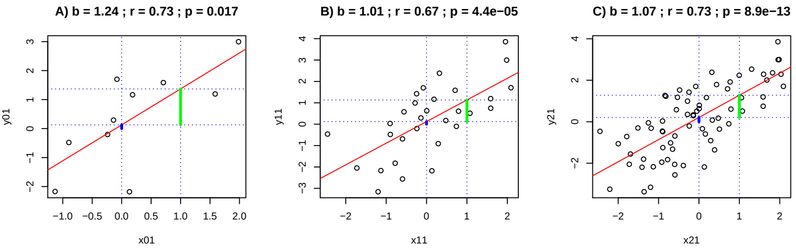
- На графике A представлено 10 точек, а на графиках B и C – 30 и 70 точек, соответственно

Статистическая значимость



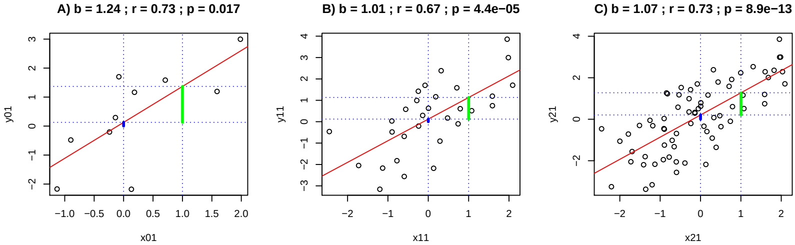
- На графике A представлено 10 точек, а на графиках B и C – 30 и 70 точек, соответственно
- Коэффициенты детерминации приблизительно равны – 0.53, 0.45, и 0.53.

Статистическая значимость



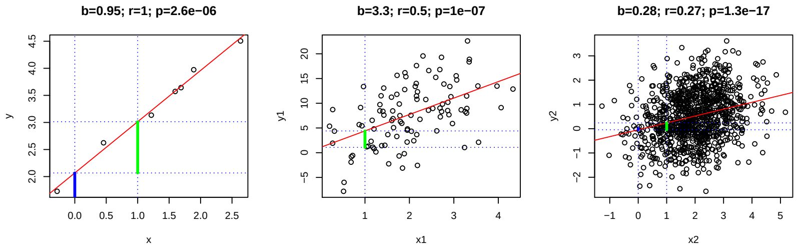
- Значения скор теста для данных A, B, и C равны $T_A^2 = n \cdot \hat{\rho}_{xy}^2 = 10 \cdot 0.53 = 5.27$; $T_B^2 = 13.63$ и $T_C^2 = 37.14$

Статистическая значимость



- Значения скор теста для данных A, B, и C равны $T_A^2 = n \cdot \hat{\rho}_{xy}^2 = 10 \cdot 0.53 = 5.27$; $T_B^2 = 13.63$ и $T_C^2 = 37.14$
- Соответствующие значения $p = 0.017$, $4.4e - 05$, и $8.9e - 13$

Так где же ассоциация сильнее?



Ответ зависит от того, как мы характеризуем ассоциацию

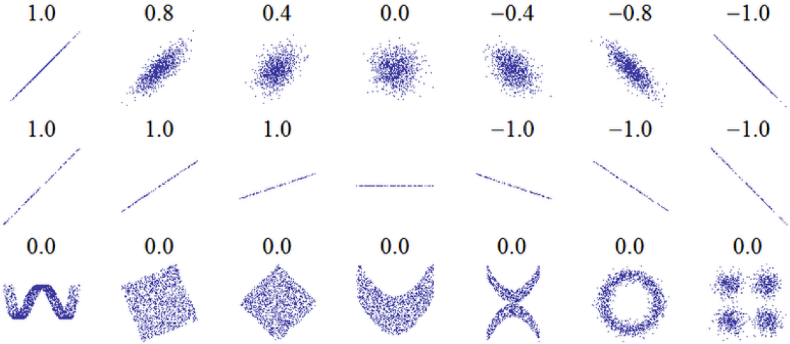
- Регрессионный коэффициент: *B*
- Корреляция: *A*
- Статистическая значимость: *C*

Заключение

Ассоциацию можно измерить несколькими взаимодополняющими способами

- Регрессионный коэффициент имееет четкую физическую интерпретацию, но зависит от шкалы как независимой, так и зависимой переменной. Поэтому сравнение возможно только для данных, представленных на одной шкале
- Коэффициенты корреляции и детерминации являются аналогами коэффициента регрессии, не зависящами от шкалы. Эти коэффициенты наиболее хорошо соответствуют интуитивному представлению о ”силе” ассоциации
- p-value отражает насколько статистически значимой является наблюдаемая ассоциация

Примечание



- Рассмотренные методы предполагают линейную зависимость между независимой и зависимой переменными
- В случае нелинейных зависимостей необходимо применение других методов

Содержание

Генетические данные

- При генетическом анализе, мы изучаем ассоциацию между независимой переменной y и генотипом g, являющемся независимой пременной
- Пусть g – SNP (single nucleotide polymorphous, однонуклеотидная замена) с двумя аллелями, A и B
- В диплоидной популяции возможны три генотипа: $\{AA, AB, BB\}$
- Различные генетические модели можно формализовать с помощью различных кодировок g

Модели с одной степенью свободы

- Оценивание одного коэффициента регрессии в рамках модели

$$y \sim \mu + \beta \cdot g,$$

где g кодируется в соответсвии с изучаемой моделью

Модели с одной степенью свободы

- Оценивание одного коэффициента регрессии в рамках модели

$$y \sim \mu + \beta \cdot g,$$

где g кодируется в соответсвии с изучаемой моделью

  - Аддитивная ("доза B"): $\{AA = 0, AB = 1, BB = 2\}$

Модели с одной степенью свободы

- Оценивание одного коэффициента регрессии в рамках модели

$$y \sim \mu + \beta \cdot g,$$

где g кодируется в соответсвии с изучаемой моделью

  - Аддитивная ("доза B"): $\{AA = 0, AB = 1, BB = 2\}$
  - Доминантный B: $\{AA = 0, AB = 1, BB = 1\}$

Модели с одной степенью свободы

- Оценивание одного коэффициента регрессии в рамках модели

$$y \sim \mu + \beta \cdot g,$$

где g кодируется в соответсвии с изучаемой моделью

  - Аддитивная ("доза B"): $\{AA = 0, AB = 1, BB = 2\}$
  - Доминантный B: $\{AA = 0, AB = 1, BB = 1\}$
  - Рецессивный B: $\{AA = 0, AB = 0, BB = 1\}$

Модели с одной степенью свободы

- Оценивание одного коэффициента регрессии в рамках модели

$$y \sim \mu + \beta \cdot g,$$

где g кодируется в соответсвии с изучаемой моделью

- Аддитивная ("доза B"): $\{AA = 0, AB = 1, BB = 2\}$
- Доминантный B: $\{AA = 0, AB = 1, BB = 1\}$
- Рецессивный B: $\{AA = 0, AB = 0, BB = 1\}$
- Сверхдоминантная (гетерозис): $\{AA = 0, AB = 1, BB = 0\}$

Генотипическая модель

- При генотипической модели, оцениваются эффекты всех трех генотипов с помощью использования двух независимых переменных

$$y \sim \mu + \beta_1 \cdot g_1 + \beta_2 \cdot g_2,$$

Генотипическая модель

- При генотипической модели, оцениваются эффекты всех трех генотипов с помощью использования двух независимых переменных

$$y \sim \mu + \beta_1 \cdot g_1 + \beta_2 \cdot g_2,$$

- g_1 and g_2 могут быть определены несколькими способами, например, g_1 можно кодировать как $\{AA = 0, AB = 1, BB = 2\}$, а g_2 – как $\{AA = 0, AB = 1, BB = 0\}$. При такой кодировке, β_1 соответствует "аддитивному эффекту B", а β_2 оценивает " доминантное отклонение"

Генотипическая модель

- При генотипической модели, оцениваются эффекты всех трех генотипов с помощью использования двух независимых переменных

$$y \sim \mu + \beta_1 \cdot g_1 + \beta_2 \cdot g_2,$$

- g_1 and g_2 могут быть определены несколькими способами, например, g_1 можно кодировать как $\{AA = 0, AB = 1, BB = 2\}$, а g_2 – как $\{AA = 0, AB = 1, BB = 0\}$. При такой кодировке, β_1 соответствует "аддитивному эффекту B", а β_2 оценивает " доминантное отклонение"
- Эта модель тестируется против нулевой модели $y \sim \mu$, тест проводится на двух степенях свободы (2 d.f.)

Заключение

- В целом, анализ ассоциации генетических данных проводится с использованием стандартных статистических методов анализа ассоциации
- Специфика этого анализа заключается в специфике независимой переменной – генотипа, который является объектом физического мира и подчиняется определенным генетическим законам