Introduction to association analysis of quantitative traits

Yurii Aulchenko

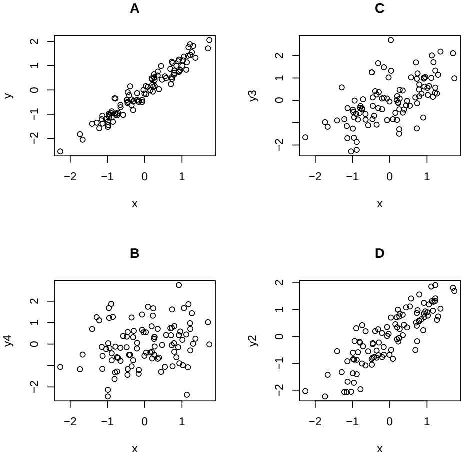
yurii [dot] aulchenko [at] gmail [dot] com

August 21, 2012

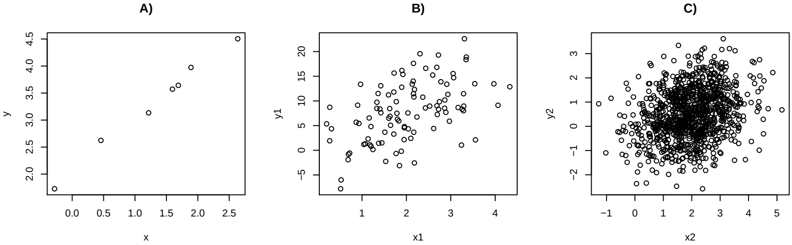
Outline

Contents

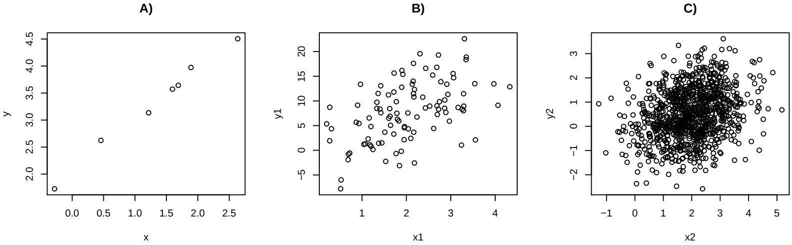
Few examples of association. Which one is stronger?



Few examples of association. Which one is stronger?

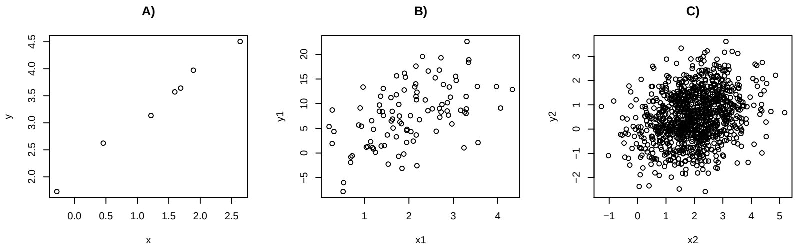


Few examples of association. Which one is stronger?



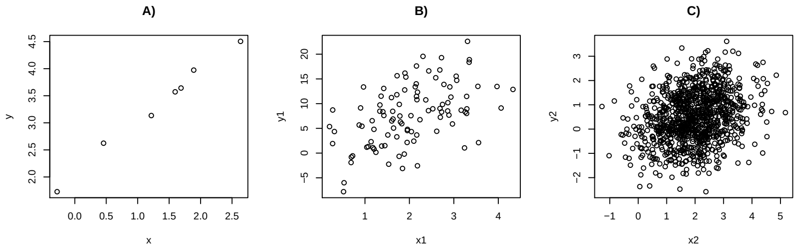
- It looks like $A > B > C$ (?)

Few examples of association. Which one is stronger?



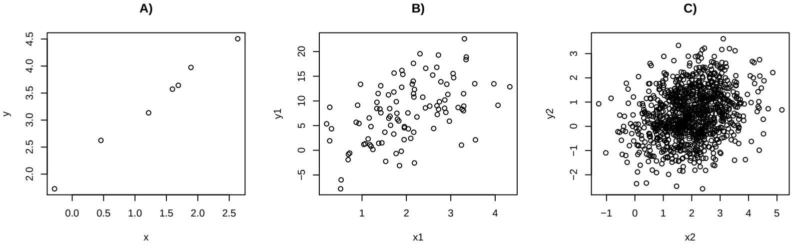
- It looks like $A > B > C$ (?)
- To give quantitative answer we need to introduce a way to characterize association between two variables

Few examples of association. Which one is stronger?



- It looks like $A > B > C$ (?)
- To give quantitative answer we need to introduce a way to characterize association between two variables
- What about using coefficient of regression of y onto x?

Few examples of association. Which one is stronger?



- It looks like $A > B > C$ (?)
- To give quantitative answer we need to introduce a way to characterize association between two variables
- What about using coefficient of regression of y onto x?
- Does everybody expect that regression coefficients $A > B > C$?

Contents

Linear regression model

- Let us denote y as "outcome" and x as "predictor" variables; let y_i and x_i are outcome and predictor values for particular sample
- Assume linear model

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

where μ is a constant (intercept), β is regression coefficient and ϵ is residual "noise"

Linear regression model

- Estimates of parameters μ and β are chosen in such a way that predicted value of outcome

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

are as close as possible to the observed y_i

- In univariate case the estimate of β can be obtained with

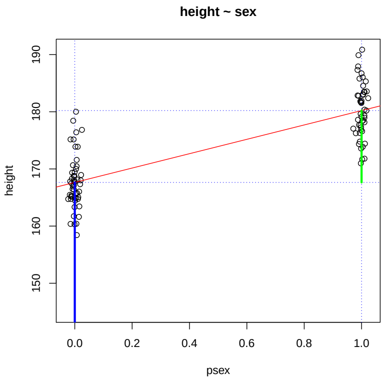
$$\hat{\beta} = \frac{Cov(x,y)}{Var(x)} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2},$$

where is $\bar{x}$ and $\bar{y}$ are mean values of x and y, respectively

Interpretation of regression coefficients

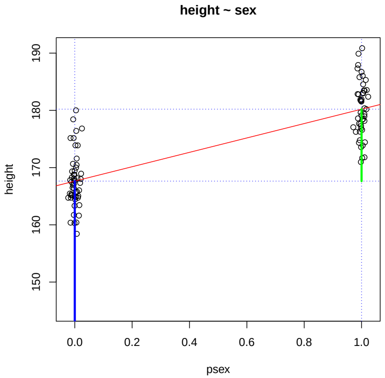
- Both intercept and regression coefficient have clear physical interpretation
- Intercept μ is expected value of y if the value of predictor x is zero
- Coefficient of regression β tells how much y change when x is changed by single unit

Example of estimation of regression coefficients



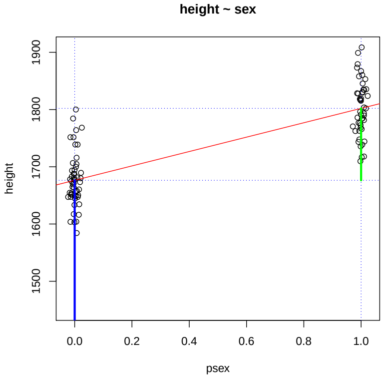
- Regression model is $y \sim \mu + \beta \cdot x$, where outcome y is height (measured in cm) and predictor x is sex (denoted as '0' for females and '1' for males)
- In a data set of 48 males and 52 females, the following estimates are obtained: $\{\hat{\mu} = 167.6, \hat{\beta} = 12.6\}$ (see figure)

Example of interpretation of regression coefficients



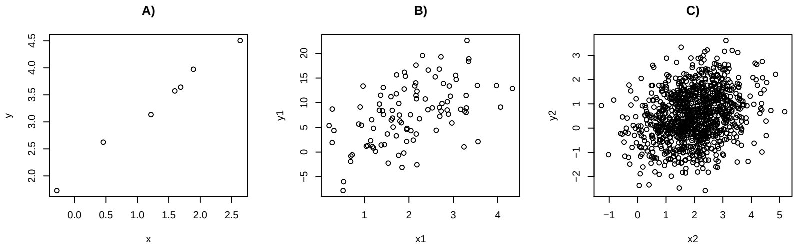
- $\hat{\mu} = 167.6$: when x is zero, expected value of outcome$y$ is 167.6. In other words, expected height of females is 167.6.
- $\hat{\beta} = 12.6$: when x changes by 1, expected value of y changes by 12.6. In other words, expected difference between male and female height is 12.6; or average height of males is $\hat{\mu} + \hat{\beta} = 180.2$

Regression coefficients are scale-dependent



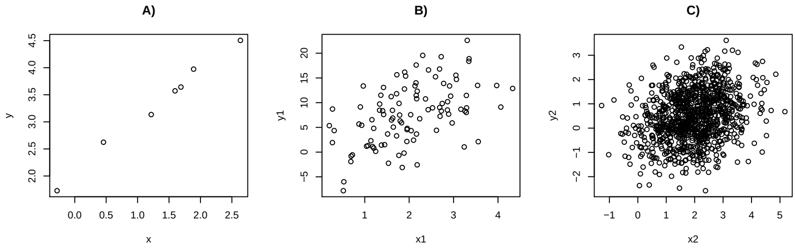
- Let height is **measured in millimeters** now
- Then the estimates are: $\{\hat{\mu} = 1676, \hat{\beta} = 126\}$
- Measuring hight in **mm** instead of **cm** $\equiv$ multiplying y by 10 $\equiv$ multiplying the estimates by 10
- But the data set is exactly the same!

Which association is stronger?



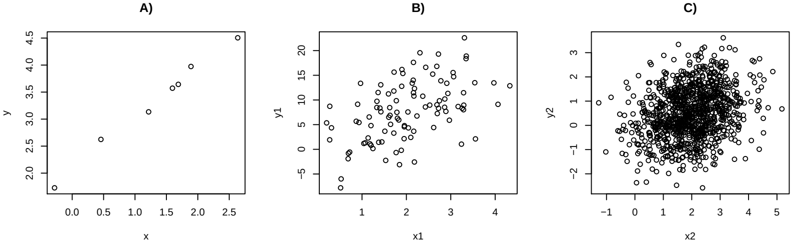
- Strength of association $A > B > C$ (?)

Which association is stronger?



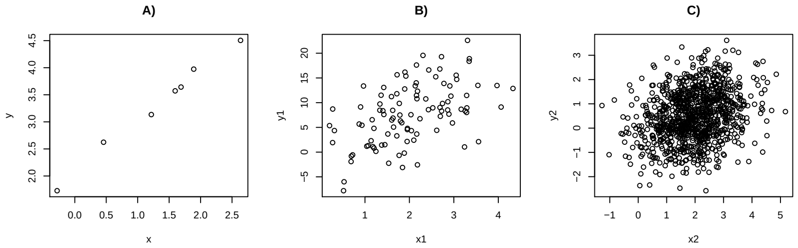
- Strength of association $A > B > C$ (?)
- Regression coefficient $A > B > C$?

Which association is stronger?



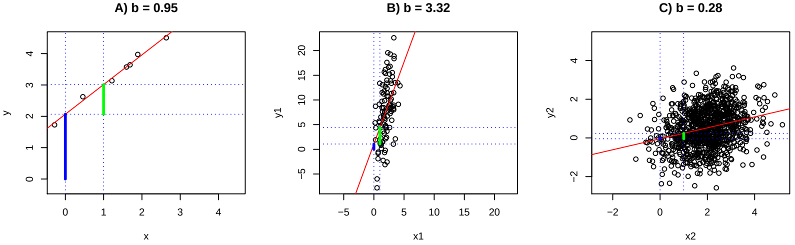
- Strength of association $A > B > C$ (?)
- Regression coefficient $A > B > C$?
- Regression coefficient may be not the best measure to characterize the strength of association because it is scale-dependent

Which association is stronger?



- Strength of association $A > B > C$ (?)
- Regression coefficient $A > B > C$?
- Regression coefficient may be not the best measure to characterize the strength of association because it is scale-dependent
- $\hat{\beta}_A = 0.95$, $\hat{\beta}_B = 3.32$ and $\hat{\beta}_C = 0.28$, so $B > A > C$

Which association is stronger?



- Strength of association $A > B > C$ (?)
- Regression coefficient $A > B > C$?
- Regression coefficient may be not the best measure to characterize the strength of association because it is scale-dependent
- $\hat{\beta}_A = 0.95$, $\hat{\beta}_B = 3.32$ and $\hat{\beta}_C = 0.28$, so $B > A > C$

How neatly y and x go together?

- We need something scale-independent!

How neatly y and x go together?

- We need something scale-independent!
- Our observation is that the regression coefficient changes with scale: the large is the variation of the outcome, the larger is the coefficient

How neatly y and x go together?

- We need something scale-independent!
- Our observation is that the regression coefficient changes with scale: the large is the variation of the outcome, the larger is the coefficient
- We can estimate how much y changes when x is changed by 1 unit with
$$\hat{\beta}_{y\sim x} = \frac{Cov(x, y)}{Var(x)}$$

How neatly y and x go together?

- We need something scale-independent!
- Our observation is that the regression coefficient changes with scale: the large is the variation of the outcome, the larger is the coefficient
- We can estimate how much y changes when x is changed by 1 unit with

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

- We can also estimate how much x (!) changes when y is changed by 1 unit with

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

Person's coefficient of correlation

- Scale independent measure of association can be obtained by "compensating" for the variance of y by use of the coefficient of correlation defined as

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

Person's coefficient of correlation

- Scale independent measure of association can be obtained by "compensating" for the variance of y by use of the coefficient of correlation defined as

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

- When
  - $\rho_{xy} = 1$, there is perfect linear dependency (as x increases, y also increases)
  - $\rho_{xy} = -1$ there is perfect reciprocal al relation (as x increases, y decreases)
  - $\rho_{xy} = 0$, there is no (linear) relation between two variables

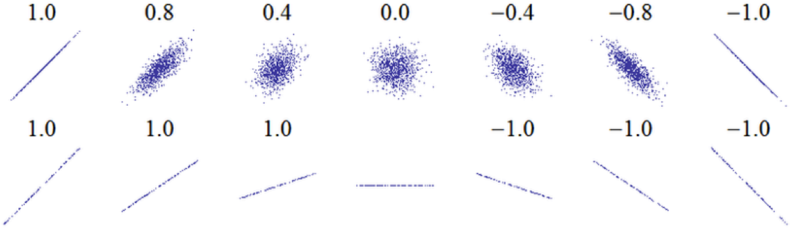
Person's coefficient of correlation

- Scale independent measure of association can be obtained by "compensating" for the variance of y by use of the coefficient of correlation defined as

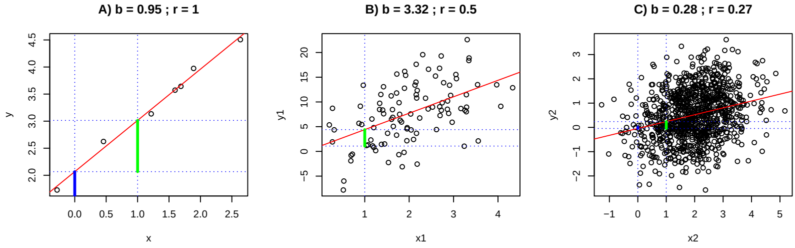
$$\rho_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2}}$$

- When
  - $\rho_{xy} = 1$, there is perfect linear dependency (as x increases, y also increases)
  - $\rho_{xy} = -1$ there is perfect reciprocal al relation (as x increases, y decreases)
  - $\rho_{xy} = 0$, there is no (linear) relation between two variables
- gives proportion of variance explained $\rho^2_{xy} = \beta_{y\sim x} \cdot \beta_{x\sim y}$, gives proportion of variance explained

Example correlations

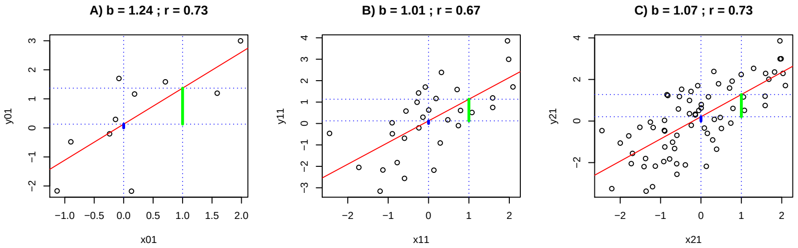


Correlations



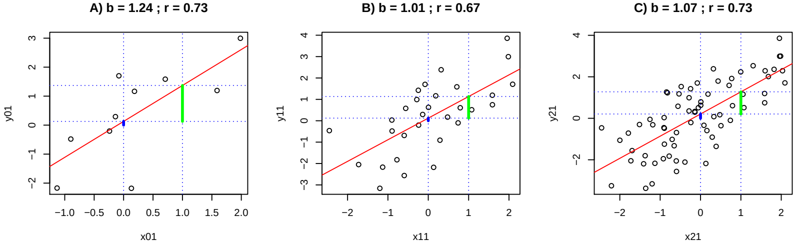
- Strength of association $A > B > C$ (?)
- Regression: $\hat{\beta}_A = 0.95$, $\hat{\beta}_B = 3.32$ and $\hat{\beta}_C = 0.28$ $(B > A > C)$
- Correlation: $\hat{\rho}_A = 1$, $\hat{\rho}_B = 0.5$ and $\hat{\rho}_C = 0.27$ $(A > B > C!)$

Other aspect of association



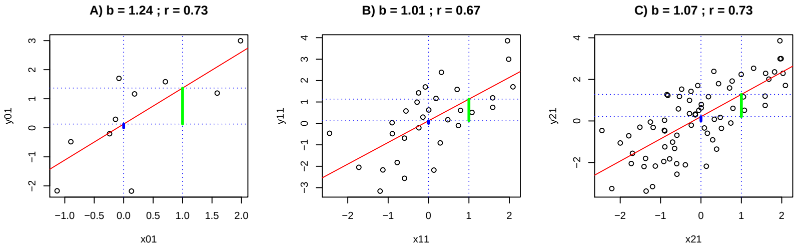
- Correlation and regression coefficients are similar between *A*, *B* and *C*
- Does that mean the same strength of association in all three panels?

Other aspect of association



- Correlation and regression coefficients are similar between A, B and C
- Does that mean the same strength of association in all three panels?
- What changes between A, B, and C?

Correlations



- There are 10 observations in panel A, 30 observations in B, and 70 observations in C. While magnitude of association is similar, amount of evidence is different
- Given the same magnitude of association, experiment with more observations provides more evidence – the observed association is less likely to appear by chance

Statistical significance

- Other way to characterize association is to ask the question "What is the chance to observe this strong (or even stronger) association by pure chance?".

Statistical significance

- Other way to characterize association is to ask the question "What is the chance to observe this strong (or even stronger) association by pure chance?".
- This chance is termed *p*-value. The lower is *p*-value, the less likely is association to appear by pure chance; consequently the statistical significance measuring our confidence is higher

The score test

- To obtain p-value, we can use the *score* test, which is defined as

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

where $\hat{\rho}^2_{xy}$ is the coefficient of determination and n is the sample size

The score test

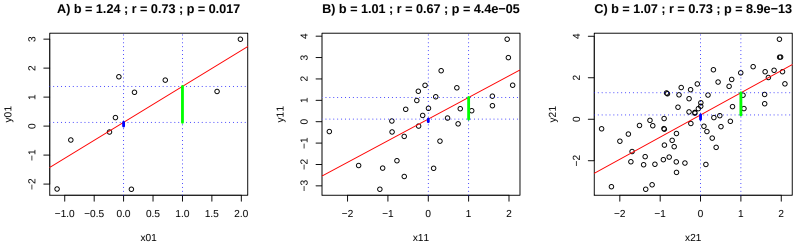
- To obtain p-value, we can use the *score* test, which is defined as

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

where $\hat{\rho}_{xy}^2$ is the coefficient of determination and n is the sample size

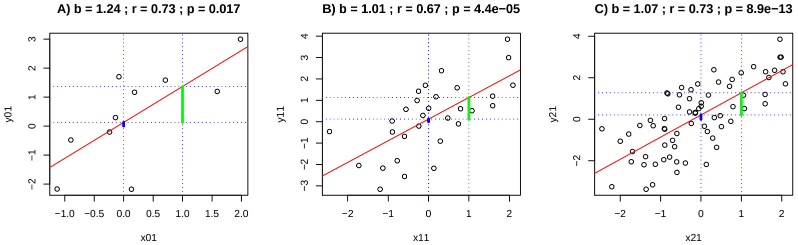
- Under the null hypothesis of no association this test is distributed as χ_1^2, so that if $T^2 > 3.84$ we can say that $p < 0.05$, etc.

Statistical significance



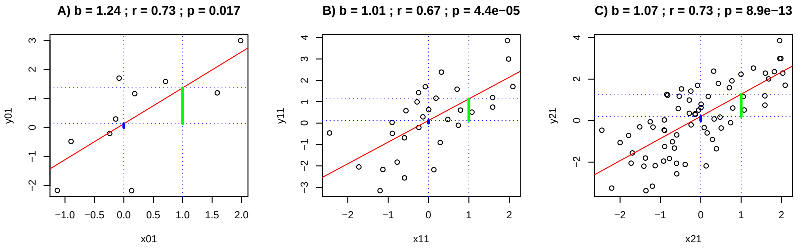
- There are 10 observations in panel *A*, 30 observations in *B*, and 70 observations in *C*.

Statistical significance



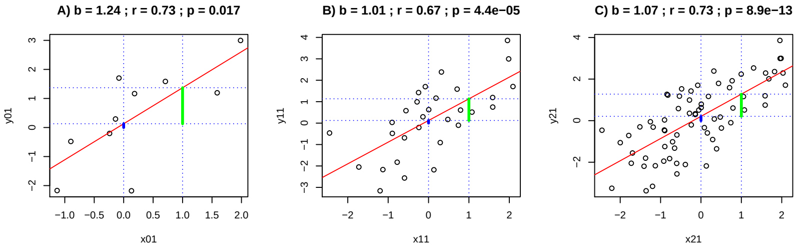
- There are 10 observations in panel A, 30 observations in B, and 70 observations in C.
- The coefficients of determination are approximately the same – 0.53, 0.45, and 0.53.

Statistical significance



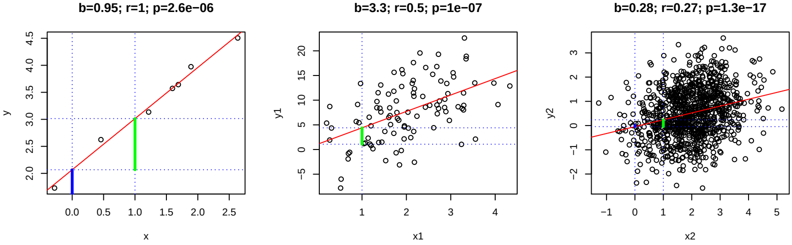
- The score test values for panels A is A, B, and C are $T_A^2 = n \cdot \hat{\rho}_{xy}^2 = 10 \cdot 0.53 = 5.27$; $T_B^2 = 13.63$ and $T_C^2 = 37.14$

Statistical significance



- The score test values for panels A is A, B, and C are $T_A^2 = n \cdot \hat{\rho}_{xy}^2 = 10 \cdot 0.53 = 5.27$; $T_B^2 = 13.63$ and $T_C^2 = 37.14$
- Resulting p-value are 0.017, $4.4e-05$, and $8.9e-13$

Which association is stronger?



The answer depends on how we characterize the association

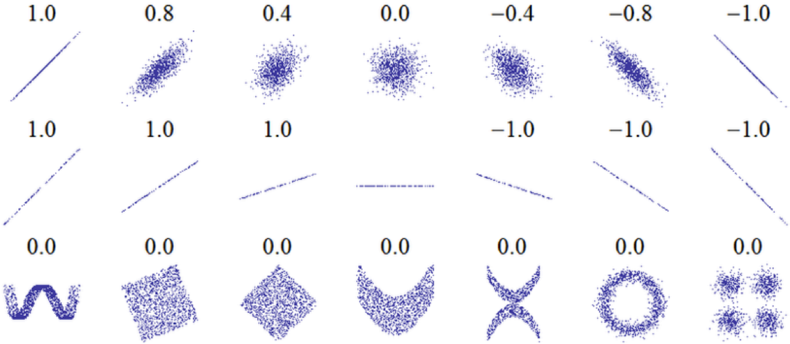
- If we use regression coefficient, then predictor $x1$ (panel B) is "the champion"
- If we use correlation or coefficient of determination, then predictor x (panel A) is "the champion"
- If we use statistical strength (p-value), then predictor $x2$ (from panel C) is "the champion"

Summary

There are several complementary ways to measure association

- Regression coefficient has clear physical interpretation and allows easy prediction. This coefficient is dependent on the scale of outcome and predictor.
- Coefficients of correlation and determination provide appealing measures of how "neatly"the outcome and the predictor go together; how "visible"is the relation
- p-value tells how much evidence are provided by the data to rule out the hypothesis of no association

Note



- Linear regression methods considered here do assume linear dependency between outcome and predictor
- While there may be a clear (non-linear) relation between two variables, methods considered here can not be used to study these

Contents

Genetic data

- When studying genetic data, we are interested in relation between outcome y and genetic predictor g
- Let g is a Single Nucleotide Polymorphism (SNP) with two alleles, A and B
- Three genotypes are possible: $\{AA, AB, BB\}$
- We can formalize different genetic models by coding g in different ways

One degree of freedom models

- Estimating single regression coefficient in the model

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

where g is coded according to different models

One degree of freedom models

- Estimating single regression coefficient in the model

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

where g is coded according to different models

  - Additive ("B allele dose"): $\{AA = 0, AB = 1, BB = 2\}$

One degree of freedom models

- Estimating single regression coefficient in the model

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

where g is coded according to different models

- Additive ("B allele dose"): $\{AA = 0, AB = 1, BB = 2\}$
- "Dominant B": $\{AA = 0, AB = 1, BB = 1\}$

One degree of freedom models

- Estimating single regression coefficient in the model

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

where g is coded according to different models

  - Additive ("B allele dose"): $\{AA = 0, AB = 1, BB = 2\}$
  - "Dominant B": $\{AA = 0, AB = 1, BB = 1\}$
  - "Recessive B": $\{AA = 0, AB = 0, BB = 1\}$

One degree of freedom models

- Estimating single regression coefficient in the model

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

where g is coded according to different models

  - Additive ("B allele dose"): $\{AA = 0, AB = 1, BB = 2\}$
  - "Dominant B": $\{AA = 0, AB = 1, BB = 1\}$
  - "Recessive B": $\{AA = 0, AB = 0, BB = 1\}$
  - Overdominant ("Heterosys") model: $\{AA = 0, AB = 1, BB = 0\}$

Genotypic model

- In genotypic model, we allow for differential effect between all three genotypes by use of two predictors

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

Genotypic model

- In genotypic model, we allow for differential effect between all three genotypes by use of two predictors

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

- g_1 and g_2 can be defined in a number of ways, for example via g_1 coded as $\{AA = 0, AB = 1, BB = 2\}$ and g_2 coded as $\{AA = 0, AB = 1, BB = 0\}$. In this case, β_1 would give "additive effect of allele B" and β_2 will estimate "dominance deviation"

Genotypic model

- In genotypic model, we allow for differential effect between all three genotypes by use of two predictors

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

- g_1 and g_2 can be defined in a number of ways, for example via g_1 coded as $\{AA = 0, AB = 1, BB = 2\}$ and g_2 coded as $\{AA = 0, AB = 1, BB = 0\}$. In this case, β_1 would give "additive effect of allele B" and β_2 will estimate "dominance deviation"
- This model is tested against the null model $y \sim \mu$, resulting in two degrees of freedom (2 d.f.) test

Summary

- In general, genetic association analysis is done using standard statistical methods
- Specifics of analysis of genetic data comes from the specifics of the independent variable of interest (the genotype), which is an real object following particular (genetic) laws