Analysis of binary traits

Yurii Aulchenko

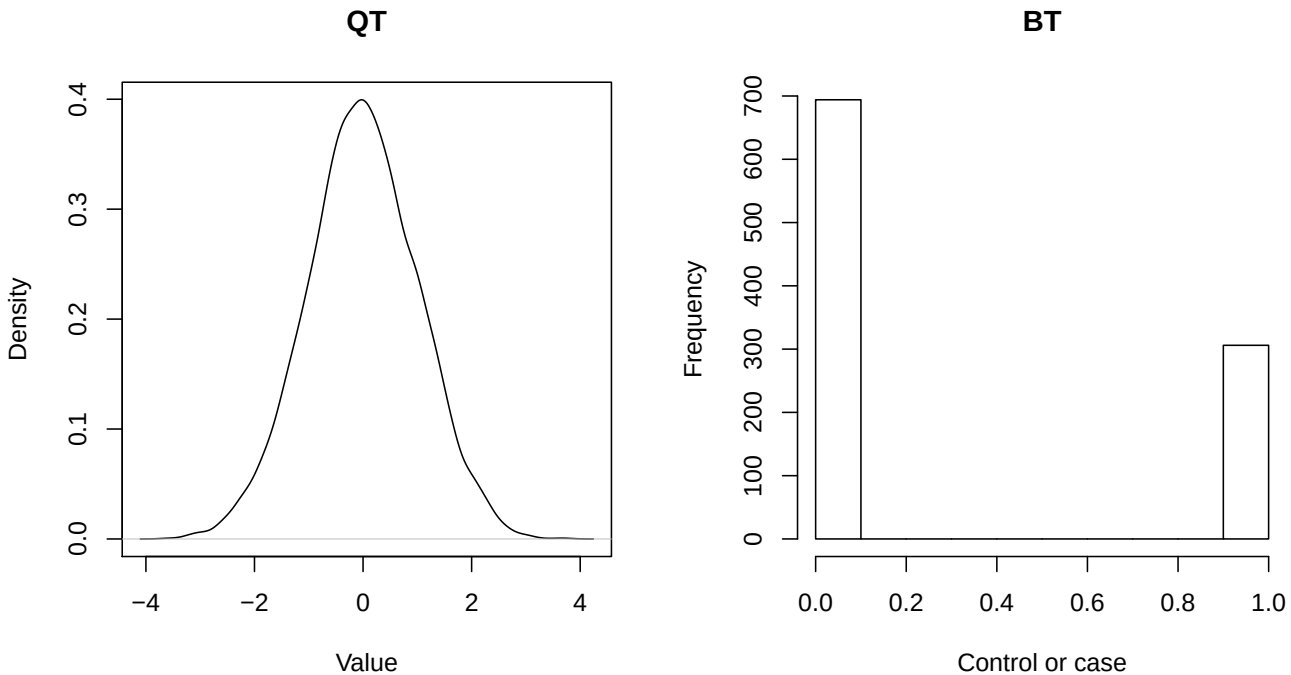
yurii [dot] aulchenko [at] gmail [dot] com

August 21, 2012

Outline

Contents

Quantitative vs. Binary



Logistic regression

- With quantitative traits, we assume linear model

$$y_i = \mu + \beta x_i + \epsilon$$

Logistic regression

- With quantitative traits, we assume linear model

$$y_i = \mu + \beta x_i + \epsilon$$

- If outcome is binary (that is y_i can be either 0 or 1) we can model expected *probability* that $y_i = 1$ using logistic function:

$$\hat{P}(y_i = 1) = \frac{1}{1 + exp\{-(\hat{\mu} + \hat{\beta} x_i)\}}$$

The same model can be expressed as

$$logit(\hat{P}(y_i = 1)) = log_e \left(\frac{\hat{P}(y_i = 1)}{1 - \hat{P}(y_i = 1)} \right) = \hat{\mu} + \hat{\beta} x_i$$

Logistic regression

- With quantitative traits, we assume linear model

$$y_i = \mu + \beta x_i + \epsilon$$

- If outcome is binary (that is y_i can be either 0 or 1) we can model expected *probability* that $y_i = 1$ using logistic function:

$$\hat{P}(y_i = 1) = \frac{1}{1 + exp\{-(\hat{\mu} + \hat{\beta} x_i)\}}$$

The same model can be expressed as

$$logit(\hat{P}(y_i = 1)) = log_e \left(\frac{\hat{P}(y_i = 1)}{1 - \hat{P}(y_i = 1)} \right) = \hat{\mu} + \hat{\beta} x_i$$

- As is the case with quantitative outcomes, the estimates of parameters μ and β are chosen in such a way as to provide maximal fit of the predicted to the observed data

Interpretation of logistic regression coeffcients

- The estimate of β are provided on logistic scale, and their physical interpretations may be difficult

Interpretation of logistic regression coeffcients

- The estimate of β are provided on logistic scale, and their physical interpretations may be difficult
- In case when the predictor is binary, Odds Ratio (OR) can be obtained from β by taking its exponent, $exp(\beta)$

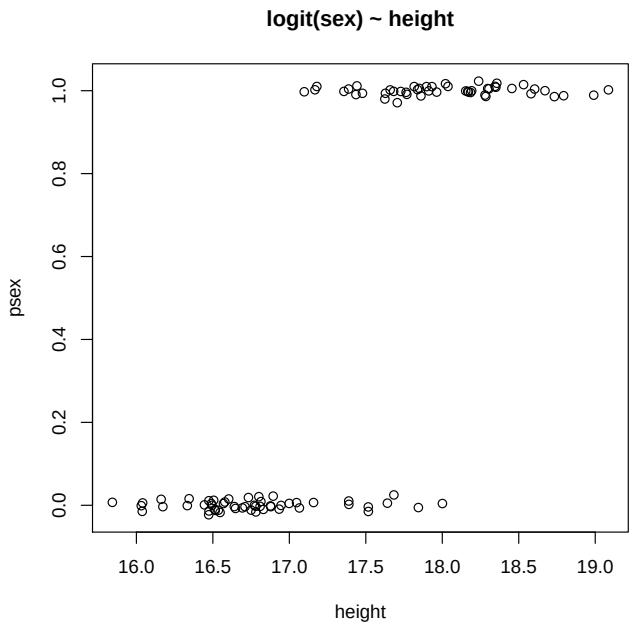
Interpretation of logistic regression coeffcients

- The estimate of β are provided on logistic scale, and their physical interpretations may be difficult
- In case when the predictor is binary, Odds Ratio (OR) can be obtained from β by taking its exponent, $exp(\beta)$
- Depending on design, OR may approximate (well or less well) the Relative Risk – how much the risk of outcome is increased when the predictor x changes by 1

Interpretation of logistic regression coeffcients

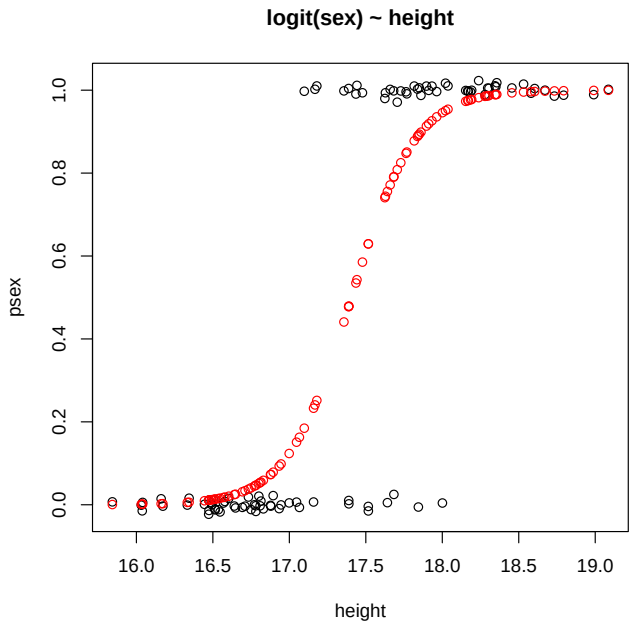
- The estimate of β are provided on logistic scale, and their physical interpretations may be difficult
- In case when the predictor is binary, Odds Ratio (OR) can be obtained from β by taking its exponent, $exp(\beta)$
- Depending on design, OR may approximate (well or less well) the Relative Risk – how much the risk of outcome is increased when the predictor x changes by 1
- For example in population-based cohort design relating some disease to the sex (0=female, 1=male), if estimate $\hat{\beta} = 0.45$, was obtained, it can be translated to $\hat{OR} = exp(0.45) = 1.49$ meaning that the risk of the disease is increased by 1.49 times in males compared to females

Example of logistic regression



- Logistic regression model is $\mathit{logit}(y) \sim \mu + \beta \cdot x$, where outcome y is sex (denoted as '0' for females and '1' for males) and predictor x is height (measured in cm)
- The following estimates are obtained: $\{\hat{\mu} = -83.7, \hat{\beta} = 4.8\}$

Example of logistic regression



- From these estimates, it is possible to predict the sex for each individual based on the height $P(\text{i is male}) = \frac{1}{1+exp(-(-83.7+4.8 \cdot height_i))}$ (red dots in the figure)

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

$$logit(y) \sim \mu + \beta \cdot g,$$

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

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

- 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

Armitage trend test

- When analyzing binary outcomes, Armitage trend test is frequently used
- This is easily performed: code g using allele dose model, and outcome as '1' for cases and '0' for controls
- Compute the coefficient of determination ρ^2 and the score test $T^2 = \rho^2 \cdot n$. This is the Armitage trend test

Contents

Summary

- Strength of association can be characterized in a number of ways

Summary

- Strength of association can be characterized in a number of ways
- For quantitative outcomes
  - **Coefficient of regression** has clear physical interpretation and allows easy prediction. This coefficient is dependent on the scale of outcome and predictor.

Summary

- Strength of association can be characterized in a number of ways
- For quantitative outcomes
  - **Coefficient of regression** 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 measure of how "neatly" the outcome and the predictor go together; how "visible" is the relation

Summary

- Strength of association can be characterized in a number of ways
- For quantitative outcomes
  - **Coefficient of regression** 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 measure of how "neatly" the outcome and the predictor go together; how "visible" is the relation
- For binary outcomes the **coefficient of regression** allows easy prediction and *some times* can be easily interpreted

Summary

- Strength of association can be characterized in a number of ways
- For quantitative outcomes
  - **Coefficient of regression** 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 measure of how "neatly" the outcome and the predictor go together; how "visible" is the relation
- For binary outcomes the **coefficient of regression** allows easy prediction and *some times* can be easily interpreted
- ***p*-value** tells how much evidence are provided by the data to rule out the hypothesis of no association