Module 11 - Population-based Association Analysis

(Fundamentals of) Statistical Genetics

Lei Sun

Department of Statistical Sciences, FAS
Division of Biostatistics, DLSPH
University of Toronto

Outline I

Chapter 7 - The Basics of (Population-based) Genetic Association Analysis

Overview of Genetic Association Analysis: QTL vs. Binary Trait

2x2 Contingency Table

- Compare two proportions using Normal test
- Likelihood Ratio Test
- Pearson test of homogeneity
- Dominant, Recessive and Allelic Models

2x3 Contingency Table

- LRT and Pearson test of homogeneity
- Genotypic model

Outline II

Alternative Approach for 2x2 Contingency Table

- Odds, and Odds Ratio
- Properties

$$\frac{odds(case|Aa)}{odds(case|aa)} = \frac{odds(Aa|case)}{odds(aa|control)}$$

- Inference: estimation and hypothesis test

Logistic Regression

- Motivation
- Interpretation of the regression parameters: link with OR
- Inference: estimation and hypothesis test
- An Example xercise 1 of Chapter 7: a 2x2 table using logistic regression
- Expanding the simple logistic regression: 2x3 table
- Additive 1 d.f. model, genotypic 2 d.f. model, dummy variables.
- A few words on generalized linear model (GLM) framework
- An Example xercise 3 of Chapter 7: a 2x3 table using logistic regression

Association Analysis Overview I

- For **a quantitative trait** Y (typically assumed to be approximately normally distributed).
  - We have learned that the mean value of Y differ only at the DSL or markers that are in LD with the DSL:

$$E(Y|G = aa) \; vs. \; E(Y|G = Aa) \; vs. \; E(Y|G = AA).$$

  - Naturally, we can use the simple linear regression models to detect association between Y and G

$$Y = \alpha + \beta G + e, e \sim N(0, \sigma^2).$$

  - Depending on how we code G, we can have 1 d.f. additive (or dominant, recessive) model or 2 d.f. genotypic model. The 1.d.f. additive model is most commonly used model.

$G =$	aa	Aa	AA
Additive	0	1	2
Dominant	0	1	1
Recessive	0	0	1
Genotypic	"0"	"1"	"2"

Association Analysis Overview II

$$Y = \alpha + \beta G + e, e \sim N(0, \sigma^2).$$

- The test of no association is equivalent to $H_0 : \beta = 0$.
- If we used 2 d.f. model, we can introduce dummy variables $D_1 = I(G = Aa)$ and $D_2 = I(G = AA)$, then $H_0 : \beta_1 = \beta_2 = 0$

$$Y = \alpha + \beta_1 D_1 + \beta_2 D_2 + e, e \sim N(0, \sigma^2).$$

- If we want to take into account any environmental factors and/or additional SNPs, then simply consider **multivariate linear regression models, possibly with interaction terms**.

$$Y = \alpha + \beta G + \gamma E + \delta GxE + e, e \sim N(0, \sigma^2).$$

Association Analysis Overview III

$$Y = \alpha + \beta G + \gamma E + \delta GxE + e, e \sim N(0, \sigma^2).$$

- From the prerequisite course, we have learned
  - Inference concerning the regression parameters: least squares estimation, hypothesis testing.
  - Interpretation of the parameters.
  - Inference from the estimated regression function: prediction.
  - Model checking: does the model fit?
  - Model selection: do we need all predictors? do we need interaction terms?
  - ....

Therefore, we will NOT spend more lecture time on this particular topic.

Association Analysis Overview IV

- For **a binary trait** Y, where $Y = 1$ denotes for being affected/cases, and $Y = 0$ for unaffected/controls.
  - We have learned that genotype (therefore allele) frequency differ between cases and controls only at the DSL or markers that are in LD with the DSL:

$$\{p_{aa,case}, p_{Aa,case}, p_{AA,case}\} \textit{ vs. } \{p_{aa,control}, p_{aa,control}, p_{aa,control}\}.$$

  - How do we detect association between Y and G?
  - Intuitively, we can **compare frequency differences between cases and controls**. We will study this.
  - But, can we **use the regression framework** to unify the analysis for QTL and binary traits, and to handle say environmental factors for binary traits? We will also study this.

Association Analysis for Binary Trait

- Table 7.1 of the Textbook: observed genotype counts at a marker under the study for r cases and s controls.

	aa	Aa	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

- Intuitively, if genotype frequency differ between cases and controls, then we can conclude that there is an associate between Y and G under the study.
- How do we proceed?

2x2 Contingency Table - Normal Test I

- Let's first consider a simplified situation where we assume AA genotype is quite rare, and we have the following data

	aa	Aa	Total
Cases	r_0	r_1	r
Controls	s_0	s_1	s
Total	n_0	n_1	n

- Parameter of interest:

$$\pi_1 = P(Aa|case), \ \ \pi_2 = P(Aa|control).$$

- Data and distribution:

$$r_1 \sim Bino(r, \pi_1), \ \ s_1 \sim Bino(s, \pi_2).$$

- Hypothesis of interest, testing the null of association between Y and G is equivalent to

$$H_0 : \pi_1 = \pi_2.$$

2x2 Contingency Table - Normal Test II

- This is essentially comparing two proportions from two (independent) binomial distributions.
- The test based on a normal approximation.
  - Independently,

$$\hat{\pi}_1 = \frac{r_1}{r} \approx N(\pi_1, \frac{\pi_1(1-\pi_1)}{r}),$$
$$\hat{\pi}_2 = \frac{s_1}{s} \approx N(\pi_2, \frac{\pi_2(1-\pi_2)}{s}).$$

  - Thus,

$$\hat{\pi}_1 - \hat{\pi}_2 \approx N(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{r} + \frac{\pi_2(1-\pi_2)}{s}).$$

  - The null hypothesis:

$$H_0 : \pi_1 = \pi_2 \equiv \pi, \text{ or } \pi_1 - \pi_2 = 0.$$

2x2 Contingency Table - Normal Test III

- Under the null hypothesis

$$\hat{\pi}_1 - \hat{\pi}_2 \approx N(0, \frac{\pi(1-\pi)}{r} + \frac{\pi(1-\pi)}{s}).$$

- Estimate $\hat{\pi}$ by pooling the case and control data together

$$\hat{\pi} = \frac{r_1 + s_1}{r + s} = \frac{n_1}{n}.$$

Note that since under the null there is no difference between cases and controls and they share a common π, we have
$n_1 = r_1 + s_1 \sim Bino(n = r + s, \pi)$, hence the above MLE.

- Therefore,

$$Var(\hat{\pi}_1 - \hat{\pi}_2) \approx \frac{\hat{\pi}(1-\hat{\pi})}{r} + \frac{\hat{\pi}(1-\hat{\pi})}{s}.$$

- Thus, our test statistic

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{r} + \frac{\hat{\pi}(1-\hat{\pi})}{s}}} \sim N(0,1).$$

2x2 Contingency Table - LRT I

- Likelihood ratio test.

$$T = 2log\frac{L(\hat{\theta})}{L(\tilde{\theta})} = 2(l(\hat{\theta}) - l(\tilde{\theta})) \sim \chi^2_r,$$

where r is the number of restrictions on θ required to define H_0.

  - Our parameters

$$\theta = (\pi_1, \pi_2).$$

  - The log (kernel) likelihood is

$$l(\pi_1, \pi_2) = \underline{r_0 log(1-\pi_1) + r_1 log(\pi_1)} + \underline{s_0 log(1-\pi_2) + s_2 log(\pi_2)}.$$

  - Under the null hypothesis $H_0 : \pi_1 = \pi_2 \equiv \pi$. (restricted),

$$l(\pi, \pi) = (r_0 + s_0)log(1-\pi) + (r_1 + r_2)log(\pi).$$

$$\tilde{\pi} = \tilde{\pi}_1 = \tilde{\pi}_2 = \frac{r_1 + s_1}{r + s} = \frac{n_1}{n}.$$

2x2 Contingency Table - LRT II

- Under the alternative hypothesis (unrestricted),

$$\hat{\pi}_1 = \frac{r_1}{r}, \quad \hat{\pi}_2 = \frac{s_1}{s}.$$

- Test statistic:

$$T = 2(l(\hat{\pi}_1, \hat{\pi}_2) - l(\tilde{\pi}_1, \tilde{\pi}_2)) \sim \chi^2_1,$$

  d.f.$= 1$ in our case because $H_0 : \pi_1 = \pi_2$.

- If we plug in our the estimate

$$T = 2\left(r_0 log(\frac{r_0}{rn_0/n}) + r_1 log(\frac{r_1}{rn_1/n}) + s_0 log(\frac{s_0}{sn_0/n}) + s_1 log(\frac{s_1}{sn_1/n})\right).$$

- Note that the above formula can be written as (as we have seen for testing one proprotion)

$$2\sum \text{observed} \times log(\frac{\text{observed}}{\text{expected}}),$$

  where expected count is under the null hypothesis of no difference cases and controls, therefore we use the marginal totals, e.g. $E(r_1) = r\tilde{\pi}_1 = r\frac{n_1}{n}$.

2x2 Contingency Table - Pearson's Homogeneity Test I

- Pearson χ^2 test of homogeneity.
  - Assume we have independent observations from I multinomial distributions, each of which has J categories, e.g. $I = 2$

	$X = 0$	$X = 1$	...	$X = J - 1$	Total
Cases	r_0	r_1	...	r_{J-1}	r
Controls	s_0	s_1	...	s_{J-1}	s
Total	n_0	n_1	...	n_{J-1}	n

  - We wish to test whether the cell probabilities of the multinomials equal, i.e. test the homogeneity of the multinomial distributions, e.g.

$$H_0 : \pi_{1j} = \pi_{2j} \equiv \pi_j, \ \ j = 0, 1, ..., J - 2.$$

2x2 Contingency Table - Pearson's Homogeneity Test II

- Test statistic:

$$T = \sum \frac{(O-E)^2}{E} \sim \chi^2_r.$$

where E is the expected count under the null hypothesis of homogeneity using pooled estimate, $\tilde{\pi}_j = \frac{n_j}{n}$.

- The degrees of freedom: $r = (I-1)(J-1)$
  $=$ number of independent counts $(I(J-1))$ $-$ the number of independent parameters estimated under the null from the data $(J-1)$.

- In our case

	aa	Aa	Total
Cases	r_0	r_1	r
Controls	s_0	s_1	s
Total	n_0	n_1	n

$$T = \frac{(r_0 - \frac{rn_0}{n})^2}{\frac{rn_0}{n}} + \frac{(r_1 - \frac{rn_1}{n})^2}{\frac{rn_1}{n}} + \frac{(s_0 - \frac{sn_0}{n})^2}{\frac{sn_0}{n}} + \frac{(s_1 - \frac{sn_1}{n})^2}{\frac{sn_1}{n}} \sim \chi^2_1.$$

2x2 Contingency Table - Dominant, Recessive, Allelic Models I

- If we want to consider a dominant model as in Table 7.2 of the Textbook, then we still have a 2x2 table and can use any of the tests above except replace r_1 with $r_1 + r_2$ etc.

	aa	Aa or AA	Total
Cases	r_0	$r_1 + r_2$	r
Controls	s_0	$s_1 + s_2$	s
Total	n_0	$n_1 + n_2$	n

- If we want to consider a recessive model, again we have a 2x2 table and all tests above directly apply as long as we keep the notations right.

	aa or Aa	AA	Total
Cases	$r_0 + r_1$	r_2	r
Controls	$s_0 + s_1$	s_2	s
Total	$n_0 + n_1$	n_2	n

2x2 Contingency Table - Dominant, Recessive, Allelic Models II

- We can also consider the allelic test (details are in Box 7.1 and Table 7.3). It's essentially a 2x2 Table analysis where the cell counts are the counts for the two types of alleles.

	a	A	Total
Cases	$2r_0 + r1$	$r_1 + 2r_2$	$2r$
Controls	$2s_0 + s_1$	$s_1 + s_2$	$2s$
Total	n_a	n_A	$2n$

- Note that we do need to assume samples are independent of each other within cases and controls and between cases and controls, so that we could assume two independent binomial distributions for the count data.

2x3 Contingency Table I

- Now, let's consider all three genotype together and consider the 2 d.f. genotypical or codominant model.

	aa	Aa	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

- We can no longer use the Z test since we have to compare more than 2 proportions. But, we can continue to use the LRT and Pearson's χ^2 tests quite straightforwardly!

2x3 Contingency Table II

- LRT
  - We now use the multinomial distribution

$$\{r_0, r_1, r_2) \sim Multinomial(r, \{\pi_{10}, \pi_{11}, (1 - \pi_{10} - \pi_{11})\})$$

$$\{s_0, s_1, s_2) \sim Multinomial(s, \{\pi_{20}, \pi_{21}, (1 - \pi_{20} - \pi_{21})\})$$

  - We have 4 parameters now

$$\theta = (\pi_{10}, \pi_{11}, \pi_{20}, \pi_{21}).$$

  - The log (kernel) likelihood is

$$l(\theta) = \underline{r_0 log(\pi_{10}) + r_1 log(\pi_{11}) + r_2 log(1 - \pi_{10} - \pi_{11})}$$

$$+\underline{s_0 log(\pi_{20}) + s_1 log(\pi_{21}) + s_2 log(1 - \pi_{20} - \pi_{21})}.$$

2x3 Contingency Table III

- Under the null hypothesis (restricted),

$$H_0 : \pi_{10} = \pi_{20} \equiv \pi_0, \ \ \pi_{11} = \pi_{21} \equiv \pi_1$$

$$l(\theta) = (r_0 + s_0)log(\pi_0) + (r_1 + s_1)log(\pi_1) + (r_2 + s_2)log(1 - \pi_0 - \pi_1)$$

$$\tilde{\pi}_0 = \tilde{\pi}_{10} = \tilde{\pi}_{20} = \frac{r_0 + s_0}{r + s} = \frac{n_0}{n}.$$
$$\tilde{\pi}_1 = \tilde{\pi}_{11} = \tilde{\pi}_{21} = \frac{r_1 + s_1}{r + s} = \frac{n_1}{n}.$$

- Under the alternative hypothesis (unrestricted),

$$\hat{\pi}_{10} = \frac{r_0}{r}, \ \ \hat{\pi}_{11} = \frac{r_1}{r},$$

$$\hat{\pi}_{20} = \frac{s_0}{s}, \ \ \hat{\pi}_{21} = \frac{s_1}{s}.$$

2x3 Contingency Table IV

- Test statistic:

$$T = 2(l(\hat{\theta}) - l(\tilde{\theta}) \sim \chi^2_2,$$

$r = 2$ in our case because we have two restrictions under the $H_0 : \pi_{10} = \pi_{20} \equiv \pi_0$ and $\pi_{11} = \pi_{21} \equiv \pi_1$.

- If we plug in our the estimate, again we have the LRT expressed as

$$2 \sum \text{observed} \times log(\frac{\text{observed}}{\text{expected}}),$$

- Specifically,

$$T = 2 \left(r_0 log(\frac{r_0}{rn_0/n}) + r_1 log(\frac{r_1}{rn_1/n}) + r_2 log(\frac{r_2}{rn_2/n}) \right.$$

$$\left. + s_0 log(\frac{s_0}{sn_0/n}) + s_1 log(\frac{s_1}{sn_1/n}) + s_2 log(\frac{s_2}{sn_2/n}) \right).$$

2x3 Contingency Table V

- Pearson χ^2 test
  - Now we have $I = 2$ and $J = 3$.
  - We wish to test whether the cell probabilities of the multinomials equal,

$$H_0 : \pi_{10} = \pi_{20} \equiv \pi_0, \text{ and } \pi_{11} = \pi_{21} \equiv \pi_1.$$

  - Test statistic:

$$T = \sum \frac{(O - E)^2}{E} \sim \chi^2_2,$$

    where E is the expected count under the null hypothesis of homogeneity using pooled estimate, $\tilde{\pi}_j = \frac{n_j}{n}$.

  - The degrees of freedom is $r = (I - 1)(J - 1) = 2$.

$$T = \frac{(r_0 - \frac{rn_0}{n})^2}{\frac{rn_0}{n}} + \frac{(r_1 - \frac{rn_1}{n})^2}{\frac{rn_1}{n}} + \frac{(r_2 - \frac{rn_2}{n})^2}{\frac{rn_2}{n}}$$
$$+ \frac{(s_0 - \frac{sn_0}{n})^2}{\frac{sn_0}{n}} + \frac{(s_1 - \frac{sn_1}{n})^2}{\frac{sn_1}{n}} + \frac{(s_2 - \frac{sn_2}{n})^2}{\frac{sn_2}{n}} \sim \chi^2_2.$$

Alternative Approach for 2x2 Contingency Table I

- Back to the simple 2x2 table.

	aa	Aa	Total
Cases	r_0	r_1	r
Controls	s_0	s_1	s
Total	n_0	n_1	n

- We might be interested

$$P(Y = 1|Aa) = P(case|Aa) \ \ vs. \ \ P(Y = 1|aa) = P(case|aa).$$

- Given the above case-control data, are the following estimates meaningful?

$$\hat{P}(case|Aa) = \frac{r_1}{n_1}, \ \ \hat{P}(case|aa) = \frac{r_0}{n_0}.$$

Alternative Approach for 2x2 Contingency Table II

- NO for a couple of reasons
  - cases are oversampled compared to the population proportion of cases
  - n_i are in fact random variables, while r and s are fixed based on the case-control sampling design.
- So, the formation on page 101 of the Textbook is NOT appropriate!
- Can we still validly compare the two proportions?
- For the two parameters, π_1 and π_2, there are many ways to measure possible differences, e.g.

$$\pi_1 - \pi_2 = 0?$$

$$\pi_1/\pi_2 = 1?$$

$$\frac{\pi_1}{1-\pi_1} / \frac{\pi_2}{1-\pi_2} = 1?$$

$$log(\frac{\pi_1}{1-\pi_1} / \frac{\pi_2}{1-\pi_2}) = 0?$$

Odds, Odds Ratio and Log OR

- Odds:

$$\frac{\pi}{1-\pi}.$$

- Logistic/Logit/Log-Odds:

$$\psi = logit(\pi) = log(\frac{\pi}{1-\pi}).$$

- Odds Ratio:

$$\frac{\pi_1}{1-\pi_1} / \frac{\pi_2}{1-\pi_2} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}.$$

- Log Odds Ratio (difference in logit)

$$\Delta = log(\frac{\pi_1}{1-\pi_1} / \frac{\pi_2}{1-\pi_2}) =$$

$$log(\frac{\pi_1}{1-\pi_1}) - log(\frac{\pi_2}{1-\pi_2}) = logit(\pi_1) - logit(\pi_2) = \psi_1 - \psi_2.$$

Properties of log Odds and log OR I

- Invariance: If we decide to reverse the roles of the binary outcomes and let π be the probability of failure rather than success, then the logit parameter simple changes the sign.

$$logit(1-\pi) = -logit(\pi).$$

- Unboundedness: this will come handy in formulating regression model for binary trait.

$$logit(\pi) = log(\frac{\pi}{1-\pi}) \in (-\infty, \infty).$$

- A difficulty with odds: odds ratios are not additive.
- When the risk is small: OR $\approx \pi_1/\pi_2$ (relative risk).
- **Most importantly**: the (log-)odds ratio can be estimated from either retrospective (e.g. case-control) or prospective studies.

$$\frac{\text{odds}(A|B)}{\text{odds}(A|\overline{B})} = \frac{\text{odds}(B|A)}{\text{odds}(B|\overline{A})}.$$

Properties of log Odds and log OR II

- Let's study it in bit more details.
  - Of interest are
$$\pi_1^* = P(case|Aa) \text{ and } \pi_2^* = P(case|aa).$$
  - But π_1^* and π_2^* themselves are NOT meaningful, so are $\pi_1^* - \pi_2^*$ or π_1^*/π_2^*.
  - Instead the meaningful quantities are
$$\pi_1 = P(Aa|case) \text{ and } \pi_2 = P(Aa|control).$$
  - But we have the following result that links the two quantities.
$$\frac{odds(case|Aa)}{odds(case|aa)} = \frac{\pi_1^*}{1-\pi_1^*} / \frac{\pi_2^*}{1-\pi_2^*}$$
$$= \frac{\pi_1}{1-\pi_1} / \frac{\pi_2}{1-\pi_2} = \frac{odds(Aa|case)}{odds(Aa|control)}$$

Properties of log Odds and log OR III

- Details of the derivation.

$$\frac{odds(case|Aa)}{odds(case|aa)} = \frac{\pi_1^*}{1-\pi_1^*} / \frac{\pi_2^*}{1-\pi_2^*} = \frac{P(case|Aa)}{1-P(case|Aa)} / \frac{P(case|aa)}{1-P(case|aa)}$$

$$= \frac{P(case|Aa)}{P(control|Aa)} / \frac{P(case|aa)}{P(control|aa)} = \frac{P(case, Aa)P(Aa)}{P(control, Aa)P(Aa)} / \frac{P(case, aa)P(aa)}{P(control, aa)P(aa)}$$

$$= \frac{P(case, Aa)}{P(control, Aa)} / \frac{P(case, aa)}{P(control, aa)}$$

$$= \frac{P(Aa|case)P(case)}{P(Aa|control)P(control)} / \frac{P(aa|case)P(case)}{P(aa|control)P(control)}$$

$$= \frac{P(Aa|case)}{P(aa|case)} / \frac{P(Aa|control)}{P(aa|control)}$$

$$= \frac{P(Aa|case)}{1-P(Aa|case)} / \frac{P(Aa|control)}{1-P(aa|control)} = \frac{\pi_1}{1-\pi_1} / \frac{\pi_2}{1-\pi_2} = \frac{odds(Aa|case)}{odds(aa|control)}$$

Odds Ratio Inference I

	aa	Aa	Total
Cases	r_0	r_1	r
Controls	s_0	s_1	s
Total	n_0	n_1	n

- We know that $\pi_1^* = P(case|Aa)$ and $\pi_2^* = P(case|aa)$ are not meaningful quantities, but $\pi_1 = P(Aa|case)$ and $\pi_2 = P(Aa|control)$ are, and

$$\frac{\pi_1^*}{1-\pi_1^*} / \frac{\pi_2^*}{1-\pi_2^*} = \frac{\pi_1}{1-\pi_1} / \frac{\pi_2}{1-\pi_2}$$

- So testing $H_0 : \Delta = log(\frac{\pi_1^*}{1-\pi_1^*} / \frac{\pi_2^*}{1-\pi_2^*}) = 0$ is equivalent to testing

$$H_0 : \Delta = log(\frac{\pi_1}{1-\pi_1} / \frac{\pi_2}{1-\pi_2}) = 0$$

- Q: why not test $H_0 : \frac{\pi_1^*}{1-\pi_1^*} / \frac{\pi_2^*}{1-\pi_2^*} = 1$?

Odds Ratio Inference II

- What is the MLE of Δ?
- Using the invariance property of MLE: if $\hat{\theta}$ is the MLE for θ, and if $g(\theta)$ is a transformation of θ, then the MLE for $g(\theta)$ is $g(\hat{\theta})$.
- Here $\theta = (\pi_1, \pi_2)$, $g(\theta) = \Delta = log(\frac{\pi_1}{1-\pi_1} / \frac{\pi_2}{1-\pi_2})$.
- We have already derived that

$$\hat{\pi}_1 = \frac{r_1}{r}, \quad \hat{\pi}_2 = \frac{s_1}{s}$$

- Thus,

$$\hat{\Delta} = log(\frac{\hat{\pi}_1}{1-\hat{\pi}_1} / \frac{\hat{\pi}_2}{1-\hat{\pi}_2}) = log(\frac{r_1 s_0}{r_0 s_1})$$

Odds Ratio Inference III

- What is the variance of of our MLE?

$$Var(\hat{\Delta}) = Var(log(\frac{\hat{\pi}_1}{1-\hat{\pi}_1} / \frac{\hat{\pi}_2}{1-\hat{\pi}_2})) = ?$$

- Recall the so called Delta method: if $X \sim N(\mu, \sigma^2)$ and $g(X)$ is a 'smooth' function of X , then

$$g(X) \sim N(g(\mu),\ \sigma^2 \cdot (g'(\mu))^2)$$

Odds Ratio Inference IV

- In the our log-odds ratio case, first consider the case group

$$X = \hat{\pi}_1 \sim N(\mu = \pi_1, \sigma^2 = \frac{\pi_1(1-\pi_1)}{r})$$

$$g(X) = g(\hat{\pi}_1) = log(\frac{\hat{\pi}_1}{1-\hat{\pi}_1})$$

$$g(\mu) = log(\frac{\pi_1}{1-\pi_1})$$

$$g'(X) = g'(\hat{\pi}) = (log(\frac{\hat{\pi}_1}{1-\hat{\pi}_1}))' = \frac{1}{\hat{\pi}_1} + \frac{1}{1-\hat{\pi}_1} = \frac{1}{\hat{\pi}_1(1-\hat{\pi}_1)}$$

$$g'(\mu) = \frac{1}{\pi_1(1-\pi_1)}$$

Odds Ratio Inference V

- Now use the Delta method, $g(X) \sim N(g(\mu),\ \sigma^2 \cdot (g'(\mu))^2)$, we have

$$g(X) = g(\hat{\pi}_1) = log(\frac{\hat{\pi}_1}{1-\hat{\pi}_1})$$

$$\sim N(log(\frac{\pi_1}{1-\pi_1}),\ \frac{\pi_1(1-\pi_1)}{r} \cdot (\frac{1}{\pi_1(1-\pi_1)})^2)$$

$$= N(log(\frac{\pi_1}{1-\pi_1}),\ \frac{1}{r\pi_1(1-\pi_1)}).$$

- Because $\hat{\pi}_2 \sim N(\pi_2, \frac{\pi_2(1-\pi_2)}{s})$ and $g(X) = g(\hat{\pi}_2)$ is the same function on $\hat{\pi}_2$, so we can follow the EXACT procedure to obtain

$$g(X) = g(\hat{\pi}_2) \sim N(log(\frac{\pi_2}{1-\pi_2}),\ \frac{1}{s\pi_2(1-\pi_2)}).$$

Odds Ratio Inference VI

- Now we have

$$g(X) = g(\hat{\pi}_1) \sim N(log(\frac{\pi_1}{1-\pi_1}), \quad \frac{1}{r\pi_1(1-\pi_1)}).$$

$$g(X) = g(\hat{\pi}_2) \sim N(log(\frac{\pi_2}{1-\pi_2}), \quad \frac{1}{s\pi_2(1-\pi_2)}).$$

- Note that $\hat{\pi}_1$ and $\hat{\pi}_2$ are independent of each other, and

$$\hat{\Delta} = log(\frac{\hat{\pi}_1}{1-\hat{\pi}_1} / \frac{\hat{\pi}_2}{1-\hat{\pi}_2}) = log(\frac{\hat{\pi}_1}{1-\hat{\pi}_1}) - log(\frac{\hat{\pi}_2}{1-\hat{\pi}_2}) = g(\hat{\pi}_1) - g(\hat{\pi}_2).$$

- Thus

$$Var(\hat{\Delta}) = Var(g(\hat{\pi}_1)) + Var(g(\hat{\pi}_2)) = \frac{1}{r\pi_1(1-\pi_1)} + \frac{1}{s\pi_2(1-\pi_2)}$$

- To obtain an estimate of this variance, we plug in the MLE of π_1 and π_2,

$$\widehat{Var(\hat{\Delta})} = \frac{1}{r_0} + \frac{1}{r_1} + \frac{1}{s_0} + \frac{1}{r_1}.$$

Odds Ratio Inference VII

- Putting everything together, we have

$$\hat{\pi}_1 = \frac{r_1}{r}, \quad \hat{\pi}_2 = \frac{s_1}{s}$$

$$\hat{\Delta} = log(\frac{\hat{\pi}_1}{1-\hat{\pi}_1} / \frac{\hat{\pi}_2}{1-\hat{\pi}_2}) = log(\frac{r_1 s_0}{r_0 s_1})$$

$$\widehat{Var(\hat{\Delta})} = \frac{1}{r_0} + \frac{1}{r_1} + \frac{1}{s_0} + \frac{1}{r_1}.$$

- In fact, we also know the (approximate) distribution of $\hat{\Delta}$ is normal:

$$\hat{\Delta} \sim N(log(\frac{\pi_1}{1-\pi_1}) - log(\frac{\pi_2}{1-\pi_2}), \quad \frac{1}{r\pi_1(1-\pi_1)} + \frac{1}{s\pi_2(1-\pi_2)}).$$

- It is now straightforward to conduct confidence interval or perform hypothesis testing of

$$H_0 : \Delta = log(\frac{\pi_1}{1-\pi_1} / \frac{\pi_2}{1-\pi_2}) = log(\frac{\pi_1}{1-\pi_1}) - log(\frac{\pi_2}{1-\pi_2}) = 0$$

Additional Considerations

- Small sample inference
  - Some counts might be small or even zero.
  - One proportion: Binomial exact test.
  - 2x2 table: Fisher's exact test (Hypergeometric distribution assuming both margins of the table, r, s, n_0, n_1 are fixed.
  - Continuity correction with large-sample methods: approximation to the exact small-sample methods, e.g.

$$\hat{\Delta} = log(\frac{(r_1 + 0.5)(s_0 + 0.5)}{(r_0 + 0.5)(s_1 + 0.5)})$$

$$\widehat{Var(\hat{\Delta})} = \frac{1}{r_0 + 0.5} + \frac{1}{r_1 + 0.5} + \frac{1}{s_0 + 0.5} + \frac{1}{r_1 + 0.5}.$$

  - Simulation-based empirical method: simulate data under the null, each time obtain the test statistic and repeat independently many times to obtain the 'background', i.e. the distribution of the test statistic under the null.

- More than 2x2 table: say how do we consider the 3 genotypes jointly?

Logistic Regression Motivation I

- Back the 2x2 contingency table again,

	aa	Aa	Total
Cases	r_0	r_1	r
Controls	s_0	s_1	s
Total	n_0	n_1	n

- How do we jointly consider 3 genotypes a SNP, or more categories for other types of markers (general two-way table)?
- How do we incorporate other covariates such as age and sex (beyond the two-way table)?

- We will consider the regression framework to describe the effects of multiple predictors Xs on a response variable Y, when Y is a binary response variable.

Logistic Regression Motivation II

- Define the response variable Y, $Y = 1$ for case and $Y = 0$ for control.
- Define the predictor variable X, $X = 1$ for genotype Aa and $X = 0$ for *aa*.
- So our data will be $(Y_i, X_i), i = 1, \ldots, n$ for n individuals in a sample.

- We are interested in whether X are Y are associated, how X influence Y, or how Y depends on the value of the predictor X.

- Recall that the traditional linear regression setting, we are interested if the mean of Y value changes for different X value, That is, how $E(Y|X) = \mu(X)$ is as a function of X, $f(X)$, e.g. the linear model:

$$E(Y|X) = \mu(X) = \alpha + \beta X.$$

- In the binary outcome case, this is

$$E(Y|X) = \mu(X) = P(Y = 1|X) = \alpha + \beta X.$$

Logistic Regression Motivation III

- Issues with this model:

$$P(Y = 1|X) = \alpha + \beta X$$

  - The **structural defect** of this model is that probability (left side) falls between 0 and 1, but a linear function (right side) takes values over entire real line.
  - We want to construct our model so that the predicted value of $P(Y = 1|X) = \mu(X)$ is bounded between 0 and 1.

  - The **interpretation** of β is

$$\beta = P(Y = 1|X = 1) - P(Y = 1|X = 0) = \alpha + \beta - \alpha$$

$$= P(Y = 1|Aa) - P(Y = 1|aa),$$

    and we know that this is a quantity that we should NOT estimate from a case-control study design.
  - We want to construct our model so that the regression coefficient β is meaningful even under case-control study design.

Logistic Regression I

- Consider logit of μ as a linear function of X.

$$\mu(X) = E(Y|X) = P(Y = 1|X)),$$

$$logit(P(Y = 1|X)) = logit(\mu(X)) = log\frac{\mu(X)}{1-\mu(X)} = \alpha + \beta X$$

$$\Longrightarrow P(Y = 1|X) = \frac{exp(\alpha + \beta X)}{1 + exp(\alpha + \beta X)},$$

$$P(Y = 0|X) = \frac{1}{1 + exp(\alpha + \beta X)}.$$

Logistic Regression II

- **What is the interpretation of the parameters, α and β?**

$$logit(\mu(X)) = log\frac{\mu(X)}{1-\mu(X)} = \alpha + \beta X.$$

- For individuals carry genotype aa, $X = 0$:

$$logit(\mu(X = 0)) = \alpha.$$

- Thus,

$$\alpha = logit(\mu(X = 0)) = log(\frac{\mu(X = 0)}{1-\mu(X = 0)}) = log(\frac{P(Y = 1|aa)}{1-P(Y = 1|aa)})$$

is the log-odds of being affected/case for genotype aa (however, its estimate estimated from a case-control study is not meaningful).

Logistic Regression III

- For individuals carry genotype Aa, $X = 1$:

$$logit(\mu(X = 1)) = \alpha + \beta.$$

- Thus,

$$\beta = logit(\mu(X = 1)) - logit(\mu(X = 0)) = log(\frac{\mu(X = 1)}{1 - \mu(X = 1)}) - log(\frac{\mu(X = 0)}{1 - \mu(X = 0)})$$

$$= log(\frac{P(Y = 1|Aa)}{1 - P(Y = 1|Aa)}) - log(\frac{P(Y = 1|aa)}{1 - P(Y = 1|aa)}) = log(\frac{\pi_1^*}{1 - \pi_1^*}) - log(\frac{\pi_2^*}{1 - \pi_2^*})$$

$$= log(\frac{P(Aa|Y = 1)}{1 - P(Aa|Y = 1)}) - log(\frac{P(Aa|Y = 0)}{1 - P(Aa|Y = 0)}) = log(\frac{\pi_1}{1 - \pi_1}) - log(\frac{\pi_2}{1 - \pi_2}) = \Delta!$$

- So, β is the log-odds ratio of being affected/case for individuals with genotype Aa ($X = 1$ copy of allele A) compared with being affected/case for individuals with genotype aa ($X = 0$ copies of allele A). which $\equiv \Delta$, the log-odds ratio of having genotype Aa among cases compared with having genotype Aa among controls.
- Because Δ can be meaningfully estimated from a case-control study so is β from the logistic regression model!

Logistic Regression Likelihood I

- Assume that the relationship between a binary trait/outcome Y and a marker/covariate X follows the logistic regression model

$$logit(P(Y=1|X)) = logit(\mu(X)) = log(\frac{\mu(X)}{1-\mu(X)}) = \alpha + \beta X.$$

- Suppose we sample n individuals from a population at random and record each individual the values of Y and X.

Individual	Response	Covariate
1	y_1	x_1
2	y_2	x_2
3	y_3	x_3
.	.	.
.	.	.
.	.	.
n	y_n	x_n

- From this sample, we wish to estimate the population parameters α and β.

Logistic Regression Likelihood II

- We note that

$$P(Y_i = 1|X_i = x_i) = \frac{exp(\alpha + \beta x_i)}{1 + exp(\alpha + \beta x_i)}$$

$$P(Y_i = 0|X_i = x_i) = 1 - P(Y_i = 1|X_i = x_i) = \frac{1}{1 + exp(\alpha + \beta x_i)}$$

- Combine both and we get

$$P(Y_i = y_i|X_i = x_i) = \frac{exp((\alpha + \beta x_i) \cdot y_i)}{1 + exp(\alpha + \beta x_i)}, \quad y_i = 0 \text{ or } 1.$$

- Since the samples are assumed to be independent of each other,

$$P(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n|X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$$

$$= \prod_{i=1}^{n} P(Y_i = y_i|X_i = x_i).$$

Logistic Regression Likelihood III

- The likelihood and log likelihood of the data in terms of $\theta = (\alpha, \beta)$ is:

$$L(\theta) = L(\alpha, \beta) = \prod_{i=1}^{n} \frac{exp((\alpha + \beta x_i) \cdot y_i)}{1 + exp(\alpha + \beta x_i)}$$

$$l(\theta) = l(\alpha, \beta) = \sum_{i=1}^{n} log \left(\frac{exp((\alpha + \beta x_i) \cdot y_i)}{1 + exp(\alpha + \beta x_i)} \right)$$

- MLE of α and β can be obtained by using the score functions. However, there is no closed-form solutions in general. Instead, alternative techniques (e.g. the Newton-Raphson method or the Fisher scoring algorithm) are used.

Logistic Regression Likelihood IV

- **In the case of one binary predictor, one can show that**

$$\hat{\alpha} = log(\frac{r_0}{s_0}),$$

$$SE(\hat{\alpha}) = \sqrt{\frac{1}{r_0} + \frac{1}{s_0}},$$

$$\hat{\beta} = log(\frac{r_1 s_0}{r_0 s_1}),$$

$$SE(\hat{\beta}) = \sqrt{\frac{1}{r_0} + \frac{1}{r_1} + \frac{1}{s_0} + \frac{1}{s_1}}.$$

- Results for β are identical to the ones derived from a 2×2 table for log OR!

MLE Derivation Details (More Advanced) I

- Note that $r_0 : Y = 1, X = 0$, $r_1 : Y = 1, X = 1$, $s_0 : Y = 0, X = 0$, $s_1 : Y = 0, X = 1$, and we can simplify the log likelihood as follows

$$l(\theta) = l(\alpha, \beta) = \sum_{i=1}^{n} log\left(\frac{exp((\alpha + \beta x_i) \cdot y_i)}{1 + exp(\alpha + \beta x_i)}\right)$$

$$= r_0 \cdot log\left(\frac{exp(\alpha)}{1 + exp(\alpha)}\right) + r_1 \cdot log\left(\frac{exp(\alpha + \beta)}{1 + exp(\alpha + \beta)}\right)$$

$$+ s_0 log \cdot \left(\frac{1}{1 + exp(\alpha)}\right) + s_1 \cdot log\left(\frac{1}{1 + exp(\alpha + \beta)}\right)$$

$$= r_0\alpha - r_0 log(1 + exp(\alpha)) + r_1(\alpha + \beta) - r_1 log(1 + exp(\alpha + \beta))$$
$$- s_0 log(1 + exp(\alpha)) - s_1 log(1 + exp(\alpha + \beta))$$

$$= r\alpha + r_1\beta - n_0 log(1 + exp(\alpha)) - n_1 log(1 + exp(\alpha + \beta))$$

MLE Derivation Details (More Advanced) II

- Obtain the score functions

$$\frac{\partial l(\theta)}{\partial \alpha} = r - n_0 \frac{exp(\alpha)}{1 + exp(\alpha)} - n_1 \frac{exp(\alpha + \beta)}{1 + exp(\alpha + \beta)}$$

$$\frac{\partial l(\theta)}{\partial \beta} = r_1 - n_1 \frac{exp(\alpha + \beta)}{1 + exp(\alpha + \beta)}$$

- Calculate the MLE

$$\frac{\partial l(\theta)}{\partial \alpha} = 0, \quad \frac{\partial l(\theta)}{\partial \beta} = 0 \Longrightarrow$$

$$r - n_0 \frac{exp(\hat{\alpha})}{1 + exp(\hat{\alpha})} - r_1 = 0 \Longrightarrow exp(\hat{\alpha}) = \frac{r_0}{s_0} \Longrightarrow \hat{\alpha} = log \frac{r_0}{s_0}$$

$$\frac{exp(\hat{\alpha} + \hat{\beta})}{1 + exp(\hat{\alpha} + \hat{\beta})} = \frac{r_1}{n_1} \Longrightarrow exp(\hat{\alpha} + \hat{\beta}) = exp(\hat{\alpha}) \cdot exp(\hat{\beta}) = \frac{r_1}{n_1 - r_1} = \frac{r_1}{s_1}$$

$$\Longrightarrow exp(\hat{\beta}) = \frac{r_1}{s_1} / \frac{r_0}{s_0} = \frac{r_1 s_0}{r_0 s_1} \Longrightarrow \hat{\beta} = log \frac{r_1 s_0}{r_0 s_1}.$$

- Variance calculation involves the second derivates and the Fisher's information.

Logistic Regression Inference Continued I

- How would we perform a formal hypothesis test of

$$H_0 : \beta = 0$$

- Wald test would be identical to the test derived for testing $\Delta = 0$

$$T = (\frac{\hat{\beta} - 0}{SE(\hat{\beta})})^2 \sim \chi^2_1.$$

$$\hat{\beta} = log(\frac{r_1 s_0}{r_0 s_1}),$$

$$SE(\hat{\beta}) = \sqrt{\frac{1}{r_0} + \frac{1}{r_1} + \frac{1}{s_0} + \frac{1}{s_1}}.$$

Logistic Regression Inference Continued II

- LRT

$$T = 2(log(L_{H_1}(\hat{\alpha}, \hat{\beta})) - log(L_{H_0}(\tilde{\alpha}, \beta = 0))) \sim \chi^2_1.$$

- $\tilde{\alpha} =?$

$$H_0 : \beta = 0,$$

$$log(\frac{\mu(X)}{1 - \mu(X)}) = \alpha.$$

$$L(\theta) = L(\alpha, 0) = \prod_{i=1}^{n} \frac{exp((\alpha) \cdot y_i)}{1 + exp(\alpha)}$$

Logistic Regression Inference Continued III

$$l(\theta) = l(\alpha, 0) = \sum_{i=1}^{n} log\left(\frac{exp((\alpha) \cdot y_i)}{1 + exp(\alpha)}\right)$$

$$= r \cdot log(\frac{exp(\alpha)}{1 + exp(\alpha)}) + s \cdot log(\frac{1}{1 + exp(\alpha)})$$

$$= r \cdot \alpha - n \cdot log(1 + exp(\alpha))$$

$$\frac{\partial l(\theta)}{\partial \alpha} = r - n \cdot \frac{exp(\alpha)}{1 + exp(\alpha)}$$

$$\frac{\partial l(\theta)}{\partial \alpha} = 0 \Longrightarrow \tilde{\alpha} = log\frac{r/n}{(n-r)/n} = log\frac{\mu}{1-\mu}.$$

- This is not surprising since r/n is a pooled estimate of $\mu = P(Y = 1)$ regardless of the value of X.

Logistic Regression Inference Continued IV

- Score test involves the score function and the Fisher's information evaluated under the null hypothesis that $\beta = 0$.

$$T = S(\tilde{\alpha}, \beta = 0)' I(\tilde{\alpha}, \beta = 0)^{-1} S(\tilde{\alpha}, \beta = 0) \sim \chi_1^2$$

- Note that the CI for OR is derived from the CI for logOR, e.g. 95% CI for OR is

$$(exp(\hat{\beta} - 1.96SE(\hat{\beta})),\ \ (exp(\hat{\beta} + 1.96SE(\hat{\beta})))$$

An Example - Exercise 1 of Chapter 7

- *The data below come from the study by Knowler et al. (1988), discussed in Chapter 3, on the association between IDDM type 2 and a haplotype from the GM system human immunoglobulin G. These data include all individuals in a sample of 4,920 Native Americans of the Pima and Papago tribes. In this example, think of the GM haplotype as just an allele at a suspected DSL.*

GM haplotype	# subjects	#(%) with IDDM
Present	293	23 (7.9)
Absent	4627	1343 (29.0)

- We can reformulate the data:

GM haplotype	affected/case (%)	unaffected/control	Total
Present, D	23 (7.9)	270	293
Absent, d	1343 (29.0)	3284	4627
Total	1366	3554	4920

- We are interested in comparing 7.9% with 29%. Note that this is not a case-control study design so these two proportions are meaningful!
- Let's use R. R codes and notes

Expanding the Simple Logistic Regression

- Now if we want to consider all three genotypes jointly.

	aa	Aa	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

- The above logistic regression can be immediately applied under the dominant or recessive disease model assumption. The only slight difference is the interpretation of α and β.

$G =$	aa	Aa	AA
Additive	0	1	2
Dominant	0	1	1
Recessive	0	0	1
Genotypic	"0"	"1"	"2"

- What about additive and genotypic model?

Simple Logistic Regression - Additive Model I

- If we consider an additive model by coding $X = 0, 1, 2$ for genotypes aa, Aa, AA to represent the number of copies of allele A, then

$$logit(P(Y = 1|X)) = logit(\mu(X)) = \alpha + \beta \cdot X$$

$$log(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)}) = \alpha + \beta \cdot X$$

- Interpretation of the parameter α: the log odds of being affected/case for the (reference/baseline) genotype group aa.

$$\alpha = logit(\mu(X = 0)) = logit(P(Y = 1|X = aa))$$

$$= log(\frac{P(Y = 1|aa)}{1 - P(Y = 1|aa)})$$

Simple Logistic Regression - Additive Model II

- Interpretation of the parameter β: the log OR of being affected/case for having 1 extra copy of allele A!

$$\beta = logit(\mu(X = x + 1)) - logit(\mu(X = x)) = \alpha + \beta(x + 1) - (\alpha + \beta x)$$

- e.g.

$$\begin{aligned}
&\beta = logit(\mu(X = 1)) - logit(\mu(X = 0)) \\
&= logit(P(Y = 1|X = Aa)) - logit(P(Y = 1|X = aa)) \\
&= log(\frac{P(Y = 1|Aa)}{1 - P(Y = 1|Aa)}) - log(\frac{P(Y = 1|aa)}{1 - P(Y = 1|aa)})
\end{aligned}$$

$$\begin{aligned}
&\beta = logit(\mu(X = 2)) - logit(\mu(X = 1)) \\
&= logit(P(Y = 1|X = AA)) - logit(P(Y = 1|X = Aa)) \\
&= log(\frac{P(Y = 1|AA)}{1 - P(Y = 1|AA)}) - log(\frac{P(Y = 1|Aa)}{1 - P(Y = 1|Aa)}).
\end{aligned}$$

- MLE of α and β in this case does not have closed-form solution. Use alternative computational algorithms, e.g. Newton-Raphson method or the Fisher scoring algorithm.

Simple Logistic Regression - Genotypic Model I

- If we do not want to restrict it to the additive model, instead we want to consider the 2 d.f. co-dominant genotypic model, how should we proceed?
- Dummy variables for covariates with > 2 levels.
  - A dummy variable: an indicator or design variable.
  - A variable with J levels/categories may be modelled using $J - 1$ dummy variables.
  - Dummy variables are often constructed so that regression coefficients provide comparisons of responses of subjects from the j_{th} category to the responses of subjects from a reference category.

Genotype	Dummy Variables	
	X_1	X_2
aa	0	0
Aa	1	0
AA	0	1

Simple Logistic Regression - Genotypic Model II

- Consider the model:

$$logit(\mu(\mathbf{X})) = \alpha + \beta_1 X_1 + \beta_2 X_2,$$

where $\mathbf{X} = (X_1, X_2)$.

- Note that
  - For individuals with genotype aa:

$$logit(\mu(X_1 = 0, X_2 = 0)) = log(\frac{P(Y = 1|aa)}{1 - P(Y = 1|aa)}) = \alpha$$

  - For individuals with genotype Aa:

$$logit(\mu(X_1 = 1, X_2 = 0)) = log(\frac{P(Y = 1|Aa)}{1 - P(Y = 1|Aa)}) = \alpha + \beta_1$$

  - For individuals with genotype AA:

$$logit(\mu(X_1 = 0, X_2 = 1)) = log(\frac{P(Y = 1|AA)}{1 - P(Y = 1|AA)}) = \alpha + \beta_2$$

Simple Logistic Regression - Genotypic Model III

- Interpretation of the parameters:
  - α: the log odds of being affected/case for (reference/baseline) genotype group aa (not meaningful from a case-control study design)

$$\alpha = logit(\mu(X_1 = 0, X_2 = 0)) = log(\frac{P(Y = 1|aa)}{1 - P(Y = 1|aa)})$$

  - β_1: the log OR of being affected/case comparing genotype Aa with genotype aa (meaningful even under a case-control study design).

$$\beta_1 = \alpha + \beta_1 - \alpha = logit(\mu(X_1 = 1, X_2 = 0)) - logit(\mu(X_1 = 0, X_2 = 0))$$

$$= log(\frac{P(Y = 1|Aa)}{1 - P(Y = 1|Aa)}) - log(\frac{P(Y = 1|aa)}{1 - P(Y = 1|aa)})$$

  - β_2: the log OR of being affected/case comparing genotype AA with genotype aa (meaningful).

$$\beta_2 = \alpha + \beta_2 - \alpha = logit(\mu(X_1 = 0, X_2 = 1)) - logit(\mu(X_1 = 0, X_2 = 0))$$

$$= log(\frac{P(Y = 1|AA)}{1 - P(Y = 1|AA)}) - log(\frac{P(Y = 1|aa)}{1 - P(Y = 1|aa)})$$

Simple Logistic Regression - Genotypic Model IV

- Estimating the parameters in the model.
  - The likelihood of the data in terms of α and βs is:

$$L(\alpha, \beta_1, \beta_2) = \prod_{i=1}^{n} \frac{exp((\alpha + \beta_1 x_{i1} + \beta_2 x_{i2}) \cdot y_i)}{1 + exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2})},$$

  - Similar approach as before: the MLE of α and β_j are obtained via iterative techniques such as the Newton-Raphson or Fisher scoring algorithms.

  - In this particular case,

$$\hat{\alpha} = log(\frac{r_0}{s_0}), \quad \hat{\beta}_1 = log(\frac{r_1 s_0}{r_0 s_1}), \quad \hat{\beta}_2 = log(\frac{r_2 s_0}{r_0 s_2}).$$

  - This result is not surprising since we allow different parameters for different comparisons: β_1 for Aa vs. aa and β_2 for AA vs. aa.
  - In the additive model, however, MLE of β involve all data.

Simple Logistic Regression - Genotypic Model V

- Hypothesis testing:

$$H_0 : \beta_1 = \beta_2 = 0$$

  - Wald test:

$$T_{wald} = (\hat{\beta}_1, \hat{\beta}_2)[Cov(\hat{\beta}_1, \hat{\beta}_2)]^{-1}(\hat{\beta}_1, \hat{\beta}_2)' \sim \chi^2_2.$$

  - Score test

$$T_{score} = S(\tilde{\alpha}, \beta_1 = 0, \beta_2 = 0)' I(\tilde{\alpha}, \beta_1 = 0, \beta_2 = 0)^{-1} S(\tilde{\alpha}, \beta_1 = 0, \beta_2 = 0) \sim \chi^2_2.$$

  - Likelihood ratio test:

$$T_{LRT} = 2log\frac{L(\hat{\theta})}{L(\tilde{\theta})} = 2(l(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2) - l(\tilde{\alpha}, \beta_1 = 0, \beta_2 = 0)) \sim \chi^2_2.$$

Simple Logistic Regression - Genotypic Model VI

- Pearson χ^2 test):

$$T_{Pearson} = \sum \frac{(O-E)^2}{E} \sim \chi^2_2,$$

where E is estimated counts under the null, i.e. using the marginal totals, r, s, n_0, n_1, n_2 and n.

Genotypic Model - Alternative Dummy Variables I

- What if we consider the following dummy variables?

Genotype	Dummy Variables	
	X_1	X_2
aa	0	0
Aa	1	0
AA	1	1

  - For individuals with genotype aa:

$$logit(\mu(X_1 = 0, X_2 = 0)) = log(\frac{P(Y = 1|aa)}{1 - P(Y = 1|aa)}) = \alpha$$

  - For individuals with genotype Aa:

$$logit(\mu(X_1 = 1, X_2 = 0)) = log(\frac{P(Y = 1|Aa)}{1 - P(Y = 1|Aa)}) = \alpha + \beta_1$$

  - For individuals with genotype AA:

$$logit(\mu(X_1 = 1, X_2 = 1)) = log(\frac{P(Y = 1|AA)}{1 - P(Y = 1|AA)}) = \alpha + \beta_1 + \beta_2$$

Genotypic Model - Alternative Dummy Variables II

- Interpretation of the parameters:
  - α: same as before, the log odds of being affected/case for (reference/baseline) genotype group aa (not meaningful from a case-control study design)

$$\alpha = logit(\mu(X_1 = 0, X_2 = 0)) = log(\frac{P(Y = 1|aa)}{1 - P(Y = 1|aa)})$$

  - β_1: the same as before, the log OR of being affected/case comparing genotype Aa with genotype aa (meaningful even under a case-control study design).

$$\beta_1 = \alpha + \beta_1 - \alpha = logit(\mu(X_1 = 1, X_2 = 0)) - logit(\mu(X_1 = 0, X_2 = 0))$$

$$= log(\frac{P(Y = 1|Aa)}{1 - P(Y = 1|Aa)}) - log(\frac{P(Y = 1|aa)}{1 - P(Y = 1|aa)})$$

  - β_2: the log OR of being affected/case comparing genotype AA with genotype Aa (meaningful).

$$\beta_2 = \alpha+\beta_1+\beta_2-(\alpha+\beta_1) = logit(\mu(X_1 = 1, X_2 = 1))-logit(\mu(X_1 = 1, X_2 = 0))$$

$$= log(\frac{P(Y = 1|AA)}{1 - P(Y = 1|AA)}) - log(\frac{P(Y = 1|Aa)}{1 - P(Y = 1|Aa)})$$

Genotypic Model - Alternative Dummy Variables III

- There is no single reference group: compare Aa with aa, and AA with Aa.
- What if we want to use Aa as the reference group, i.e.. compare aa with Aa, and AA with Aa?

Genotype	Dummy Variables	
	X_1	X_2
aa	1	0
Aa	0	0
AA	0	1

- What if we want to use AA as the reference group, i.e. compare aa with AA, and Aa with AA?

Genotype	Dummy Variables	
	X_1	X_2
aa	1	0
Aa	0	1
AA	0	0

Additional Considerations I

- Which model is better: additive vs. genotypic model?
- Considering covariates: interaction, model selection etc.

- The textbook also considered
  - Trend test and its connection with the tests discussed so far.
  - Sample size and power calculation.

Additional Considerations II

- The general Generalized Linear Model (GLM) framework:

$$g(E(Y|X)) = g(\mu(X)) = \alpha + \beta X,$$

where $g(.)$ is the link function.

  - If Y is Normal, often use 'identity' link:

$$g(E(Y|X))) = E(Y|X) = \alpha + \beta X.$$

  - If Y is Binomial as in the case-control study case, often use the logistic link:

$$E(Y|X)) = P(Y = 1|X),$$

$$g(E(Y|X)) = \frac{P(Y = 1|X)}{1 - P(Y = 1|X)} = \alpha + \beta X.$$

  - If Y is Poisson, often use the log link:

$$g(E(Y|X)) = log(E(Y|X)) = \alpha + \beta X.$$

  - etc.

An Numerical Example - Exercise 3 of Chapter 7

- *A case/control study has been conducted and a SNP genotyped. Compute the odds-ratios for the table below and the corresponding confidence intervals. Compute tests for all 3 modes of inheritance (large sample) discussed in the chapter. Discuss the results in terms of plausibility of a model.*

	aa	Aa	AA	Total
Cases	$r_0 = 500$	$r_1 = 350$	$r_2 = 120$	$r = 970$
Controls	$s_0 = 521$	$s_1 = 270$	$s_2 = 130$	$s = 921$
Total	$n_0 = 1021$	$n_1 = 620$	$n_2 = 250$	$n = 1891$

- R codes and related notes.

Exercises

- Chapter 7 Exercise 1

  We have done this in the class, but it's important that you repeat the exercise yourself. You are expected to be able to do the calculation by hand (for comparing two proportions and OR inference of 2x2 table and logistic regression inference involving one predictor with 2 levels), as well as understand R output).
- Chapter 7 Exercise 2(a) and 2(b).
- Chapter 7 Exercise 3

  Again, we have done this in the class, but it's important that you repeat the exercise yourself. Also work on the allelic test (Box 7.1 of the Textbook) which is essentially a 2x2 table with counts representing the number of alleles observed.
- Using the data in Table 7.11 and following the lecture notes and the R codes provided in the notes, repeat the tests yourself. You don't need to do for all SNPs, but it's important to fully understand the procedures including how to run and understand R outputs.

What's Next

- Chapter 8 - Population Substructure in Association Studies
- Chapter 9 - Association Analysis in Family Designs