Quantitative Genomics and Genetics
BTRY 4830/6830; PBSB.5201.03

Lecture 23-24: Logistic Regression I & II

Jason Mezey
jgm45@cornell.edu
May 6, 2021 (Th) 8:05-9:20 (+9:20-?)
Announcements

• Today’s lecture will be two lecture (23 & 24) - I realize some / many will need to leave at 9:20 (I will break for a moment at 9:20 and then continue) I WILL POST A RECORDING OF THE ENTIRE LECTURE ASAP (=before computer lab today so you can review)

• Your Final has been set (!!)
  • Take home and same structure as midterm (= take home, open book,YOU MAY NOT COMMUNICATE WITH ANYONE IN ANY WAY ABOUT THE EXAM ONCE THE EXAM HAS STARTED)

• Dates: Available Tues. May 18 and Due 11:59PM Tues. May 25

• You will be required to do a linear regression analysis of GWAS with and without covariates AND a logistic regression analysis of GWAS with and without covariates

• If you prepare your code ahead of time, the exam should take you ~6 hours total

• Note that all next week is OPTIONAL (=lectures are Bayesian Statistics + some additional optional recored lectures) + Computer Lab will be exam review (!!) and the main Bayesian algorithm (MCMC)
Summary of lecture 23-24

- Today: Logistic Regression!
Conceptual Overview

Genetic System

Does A1 → A2 affect Y?

Sample or experimental pop

Measured individuals (genotype, phenotype)

Reject / DNR

Regression model

Model params F-test

Pr(Y|X)
Linear regression review

• So far, we have considered a linear regression is a reasonable model for the relationship between genotype and phenotype (where this implicitly assumes a normal error provides a reasonable approximation of the phenotype distribution given the genotype):

\[ Y = \beta_\mu + X_a \beta_a + X_d \beta_d + \epsilon \quad \epsilon \sim N(0, \sigma_\epsilon^2) \]
Case / Control Phenotypes I

• While a linear regression may provide a reasonable model for many phenotypes, we are commonly interested in analyzing phenotypes where this is NOT a good model.

• As an example, we are often in situations where we are interested in identifying causal polymorphisms (loci) that contribute to the risk for developing a disease, e.g. heart disease, diabetes, etc.

• In this case, the phenotype we are measuring is often “has disease” or “does not have disease” or more precisely “case” or “control”.

• Recall that such phenotypes are properties of measured individuals and therefore elements of a sample space, such that we can define a random variable such as $Y(\text{case}) = 1$ and $Y(\text{control}) = 0$. 

Case / Control Phenotypes II

Let’s contrast the situation, let’s contrast data we might model with a linear regression model versus case / control data:
Case / Control Phenotypes II

- Let’s contrast the situation, let’s contrast data we might model with a linear regression model versus case / control data:
Let’s contrast the situation, let’s contrast data we might model with a linear regression model versus case / control data:
Logistic regression I

- Instead, we’re going to consider a logistic regression model
Logistic regression II

• It may not be immediately obvious why we choose regression “line” function of this “shape”

• The reason is mathematical convenience, i.e. this function can be considered (along with linear regression) within a broader class of models called Generalized Linear Models (GLM) which we will discuss next lecture

• However, beyond a few differences (the error term and the regression function) we will see that the structure and out approach to inference is the same with this model!
Logistic regression III

- To begin, let’s consider the structure of a regression model:
  \[ Y = \text{logistic}(\beta_\mu + X_a\beta_a + X_d\beta_d) + \epsilon_l \]

- We code the “X’s” the same (!!) although a major difference here is the “logistic” function as yet undefined.

- However, the expected value of \( Y \) has the same structure as we have seen before in a regression:
  \[ E(Y_i|X_i) = \text{logistic}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d) \]

- We can similarly write for a population using matrix notation (where the \( X \) matrix has the same form as we have been considering!):
  \[ E(Y|X) = \text{logistic}(X\beta) \]

- In fact the two major differences are in the form of the error and the logistic function.
Logistic regression: error term I

- Recall that for a linear regression, the error term accounted for the difference between each point and the expected value (the linear regression line), which we assume follow a normal, but for a logistic regression, we have the same case but the value has to make up the value to either 0 or 1 (what distribution is this?):

![Graphs showing linear and logistic regression](image)
Logistic regression: error term II

- For the error on an individual $i$, we therefore have to construct an error that takes either the value of “1” or “0” depending on the value of the expected value of the genotype

  - For $Y = 0$

    $$\epsilon_i = -E(Y_i|X_i) = -E(Y|A_iA_j) = -\text{logistic}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)$$

  - For $Y = 1$

    $$\epsilon_i = 1 - E(Y_i|X_i) = 1 - E(Y|A_iA_j) = 1 - \text{logistic}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)$$
Logistic regression: error term II

• For the error on an individual $i$, we therefore have to construct an error that takes either the value of “1” or “0” depending on the value of the expected value of the genotype

  • For $Y = 0$
    \[
    \epsilon_i = -E(Y_i|X_i) = -E(Y|A_iA_j) = -\text{logistic}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)
    \]

  • For $Y = 1$
    \[
    \epsilon_i = 1 - E(Y_i|X_i) = 1 - E(Y|A_iA_j) = 1 - \text{logistic}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)
    \]

• For a distribution that takes two such values, a reasonable distribution is therefore the Bernoulli distribution with the following parameter

  \[
  \epsilon_i = Z - E(Y_i|X_i)
  \]
Logistic regression: error term II

- For the error on an individual $i$, we therefore have to construct an error that takes either the value of “1” or “0” depending on the value of the expected value of the genotype

  - For $Y = 0$
    \[
    \epsilon_i = -E(Y_i|X_i) = -E(Y|A_iA_j) = -logistic(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)
    \]

  - For $Y = 1$
    \[
    \epsilon_i = 1 - E(Y_i|X_i) = 1 - E(Y|A_iA_j) = 1 - logistic(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)
    \]

- For a distribution that takes two such values, a reasonable distribution is therefore the Bernoulli distribution with the following parameter

  \[
  \epsilon_i = Z - E(Y_i|X_i)
  \]

  \[
  Pr(Z) \sim bern(p) \quad p = logistic(\beta_\mu + X_{a}\beta_a + X_{d}\beta_d)
  \]
Logistic regression: error term III

- This may look complicated at first glance but the intuition is relatively simple.

- If the logistic regression line is near zero, the probability distribution of the error term is set up to make the probability of $Y$ being zero greater than being one (and vice versa for the regression line near one!):

$$
\epsilon_i = Z - E(Y_i|X_i)
$$

$$
Pr(Z) \sim bern(p)
$$

$$
p = logistic(\beta_\mu + X_a\beta_a + X_d\beta_d)
$$
The error term I

- Recall that the error term is either the negative of \( E(Y_i \mid X_i) \) when \( Y_i \) is zero and \( 1 - E(Y_i \mid X_i) \) when \( Y_i \) is one:

\[
\epsilon_i \mid (Y_i = 0) = - E(Y_i \mid X_i) \quad \epsilon_i \mid (Y_i = 1) = 1 - E(Y_i \mid X_i)
\]

- For the entire distribution of the population, recall that

\[
Pr(\epsilon_i) \sim bern(p \mid X) - E(Y \mid X)
\]

\[
p = E(Y \mid X)
\]

For example:

\[
\epsilon_i = -0.1 \quad \epsilon_i = 0.9 \quad p = 0.1
\]
The error term II

- Recall that the error term is either the negative of $E(Y_i \mid X_i)$ when $Y_i$ is zero and $1 - E(Y_i \mid X_i)$ when $Y_i$ is one:

$$
\epsilon_i \mid (Y_i = 0) = -E(Y_i \mid X_i) \quad \epsilon_i \mid (Y_i = 1) = 1 - E(Y_i \mid X_i)
$$

- For the entire distribution of the population, recall that

$$
Pr(\epsilon_i) \sim bern(p \mid X) - E(Y \mid X)
$$

$$
p = E(Y \mid X)
$$

For example:

$$
\epsilon_i = -0.6 \quad \epsilon_i = 0.4
$$

$$
p = 0.6
$$
The error term III

• Recall that the error term is either the negative of $E(Y_i \mid X_i)$ when $Y_i$ is zero and $1 - E(Y_i \mid X_i)$ when $Y_i$ is one:

$$
\epsilon_i | (Y_i = 0) = -E(Y_i \mid X_i) \quad \epsilon_i | (Y_i = 1) = 1 - E(Y_i \mid X_i)
$$

• For the entire distribution of the population, recall that

$$
Pr(\epsilon_i) \sim \text{bern}(p \mid X) - E(Y \mid X)
$$

$$
p = E(Y \mid X)
$$

For example:

$$
\epsilon_i = -0.9 \quad \epsilon_i = 0.1
$$

$$
p = 0.9
$$
Logistic regression: link function

- Next, we have to consider the function for the regression line of a logistic regression (remember below we are plotting just versus Xa but this really is a plot versus Xa AND Xd!!):

\[ E(Y_i|X_i) = \text{logistic}(\beta_\mu + X_i,a,\beta_a + X_i,d,\beta_d) \]

\[ E(Y_i|X_i) = \frac{e^{\beta_\mu + X_i,a,\beta_a + X_i,d,\beta_d}}{1 + e^{\beta_\mu + X_i,a,\beta_a + X_i,d,\beta_d}} \]
Calculating the components of an individual $i$

- For example, say we have an individual $i$ that has genotype $A1A1$ and phenotype $Y_i = 0$
- We know $X_a = -1$ and $X_d = -1$
- Say we also know that for the population, the true parameters (which we will not know in practice! We need to infer them!) are:
  \[ \beta_\mu = 0.2 \quad \beta_a = 2.2 \quad \beta_d = 0.2 \]
- We can then calculate the $E(Y_i|X_i)$ and the error term for $i$:
  
  \[
  Y_i = \frac{e^{\beta_\mu + x_i,a\beta_a + x_i,d\beta_d}}{1 + e^{\beta_\mu + x_i,a\beta_a + x_i,d\beta_d}} + \epsilon_i
  \]
  
  \[
  0 = \frac{e^{0.2+(-1)2.2+(-1)0.2}}{1 + e^{0.2+(-1)2.2+(-1)0.2}} + \epsilon_i
  \]
  
  \[0 = 0.1 - 0.1\]
Calculating the components of an individual II

- For example, say we have an individual i that has genotype A1A1 and phenotype Yi = 1

- We know Xa = -1 andXd = -1

- Say we also know that for the population, the true parameters (which we will not know in practice! We need to infer them!) are:

  \[ \beta_\mu = 0.2 \quad \beta_a = 2.2 \quad \beta_d = 0.2 \]

- We can then calculate the E(Yi|Xi) and the error term for i:

  \[
  Y_i = \frac{e^{\beta_\mu + x_i,a \beta_a + x_i,d \beta_d}}{1 + e^{\beta_\mu + x_i,a \beta_a + x_i,d \beta_d}} + \epsilon_i
  \]

  \[
  1 = \frac{e^{0.2 + (-1)2.2 + (-1)0.2}}{1 + e^{0.2 + (-1)2.2 + (-1)0.2}} + \epsilon_i
  \]

  \[
  1 = 0.1 + 0.9
  \]
Calculating the components of an individual III

- For example, say we have an individual i that has genotype A1A2 and phenotype Yi = 0
- We know Xa = 0 andXd = 1
- Say we also know that for the population, the true parameters (which we will not know in practice! We need to infer them!) are:
  \[ \beta_\mu = 0.2 \quad \beta_a = 2.2 \quad \beta_d = 0.2 \]
- We can then calculate the E(Yi|Xi) and the error term for i:

\[
Y_i = \frac{e^{\beta_\mu + x_i,a \beta_a + x_i,d \beta_d}}{1 + e^{\beta_\mu + x_i,a \beta_a + x_i,d \beta_d}} + \epsilon_i
\]
\[
0 = \frac{e^{0.2+(0)2.2+(1)0.2}}{1 + e^{0.2+(0)2.2+(1)0.2}} + \epsilon_i
\]

\[ 0 = 0.6 - 0.6 \]
Calculating the components of an individual IV

- For example, say we have an individual i that has genotype A2A2 and phenotype Yi = 0

- We know Xa = 1 andXd = -1

- Say we also know that for the population, the true parameters (which we will not know in practice! We need to infer them!) are:

  \[ \beta_\mu = 0.2 \quad \beta_a = 2.2 \quad \beta_d = 0.2 \]

- We can then calculate the E(Yi|Xi) and the error term for i:

\[
Y_i = \frac{e^{\beta_\mu + x_{i,d}\beta_d + x_{i,a}\beta_a + x_{i,l}\beta_l}}{1 + e^{\beta_\mu + x_{i,d}\beta_d + x_{i,a}\beta_a + x_{i,l}\beta_l}} + \epsilon_i
\]

\[
0 = \frac{e^{0.2 + (1)2.2 + (0)0.2}}{1 + e^{0.2 + (1)2.2 + (0)0.2}} + \epsilon_i
\]

\[
0 = 0.9 - 0.9
\]
Notation

- Remember that while we are plotting this versus just Xa, the true plot is versus BOTH Xa and Xd (harder to see what is going on).
- For an entire sample, we can use matrix notation as follows:

\[
E(Y | X) = \gamma^{-1}(X \beta) = \frac{e^{X \beta}}{1 + e^{X \beta}} = \frac{1}{1 + e^{-X \beta}}
\]

\[
E(y | x) = \gamma^{-1}(x \beta) = \begin{bmatrix}
\frac{e^{\beta \mu + x_1, a \beta a + x_1, d \beta d}}{1 + e^{\beta \mu + x_1, a \beta a + x_1, d \beta d}} \\
\vdots \\
\frac{e^{\beta \mu + x_n, a \beta a + x_n, d \beta d}}{1 + e^{\beta \mu + x_n, a \beta a + x_n, d \beta d}}
\end{bmatrix}
\]
Inference

• Recall that our goal with using logistic regression was to model the probability distribution of a case / control phenotype when there is a causal polymorphism.

• To use this for a GWAS, we need to test the null hypothesis that a genotype is not a causal polymorphism (or more accurately that the genetic marker we are testing is not in LD with a causal polymorphism!):

\[
\begin{align*}
\beta_{\mu} &= c \quad \beta_a = 0 \quad \beta_d = 0 \\
H_0 : \beta_a &= 0 \cap \beta_d = 0
\end{align*}
\]

• To assess this null hypothesis, we will use the same approach as in linear regression, i.e. we will construct a LRT = likelihood ratio test (recall that an F-test is an LRT!)

• We will need MLE for the parameters of the logistic regression for the LRT.
MLE of logistic regression parameters

- Recall that an MLE is simply a statistic (a function that takes the sample as an input and outputs the estimate of the parameters)!

- In this case, we want to construct the following MLE:

\[ MLE(\hat{\beta}) = MLE(\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d) \]

- To do this, we need to maximize the log-likelihood function for the logistic regression, which has the following form (sample size n):

\[
l(\beta) = \sum_{i=1}^{n} [y_i \ln(\gamma^{-1}(\beta_\mu + x_{i,a} \beta_a + x_{i,d} \beta_d)) + (1 - y_i) \ln(1 - \gamma^{-1}(\beta_\mu + x_{i,a} \beta_a + x_{i,d} \beta_d))]
\]

- Unlike the case of linear regression, where we had a “closed-form” equation that allows us to plug in the Y's and X's and returns the beta values that maximize the log-likelihood, there is no such simple equation for a logistic regression

- We will therefore need an \textit{algorithm} to calculate the MLE
Algorithm Basics

- **algorithm** - a sequence of instructions for taking an input and producing an output

- We often use algorithms in estimation of parameters where the structure of the estimation equation (e.g., the log-likelihood) is so complicated that we cannot
  - Derive a simple (closed) form equation for the estimator
  - Cannot easily determine the value the estimator should take by other means (e.g., by graphical visualization)

- We will use algorithms to “search” for the parameter values that correspond to the estimator of interest

- Algorithms are not guaranteed to produce the correct value of the estimator (!!), because the algorithm may “converge” (=return) the wrong answer (e.g., converges to a “local” maximum or does not converge!) and because the compute time to converge to exactly the same answer is impractical for applications
IRLS algorithm I

- For logistic regression (and GLM’s in general!) we will construct an algorithm to find the parameters that correspond to the maximum of the log-likelihood:

\[
l(\beta) = \sum_{i=1}^{n} \left[ y_i ln(\gamma^{-1}(\beta_\mu + x_{i,a} \beta_a + x_{i,d} \beta_d)) + (1 - y_i) ln(1 - \gamma^{-1}(\beta_\mu + x_{i,a} \beta_a + x_{i,d} \beta_d)) \right]
\]

- For logistic regression (and GLM’s in general!) we will construct an Iterative Re-weighted Least Squares (IRLS) algorithm, which has the following structure:

1. Choose starting values for the \( \beta \)'s. Since we have a vector of three \( \beta \)'s in our case, we assign these numbers and call the resulting vector \( \beta^{[0]} \).

2. Using the re-weighting equation (described next slide), update the \( \beta^{[t]} \) vector.

3. At each step \( t > 0 \) check if \( \beta^{[t+1]} \approx \beta^{[t]} \) (i.e. if these are approximately equal) using an appropriate function. If the value is below a defined threshold, stop. If not, repeat steps 2,3.
Step 1: IRLS algorithm

1. Choose starting values for the $\beta$’s. Since we have a vector of three $\beta$’s in our case, we assign these numbers and call the resulting vector $\beta^{[0]}$.

- These are simply values of the vector that we assign (!!)
- In one sense, these can be anything we want (!!) although for algorithms in general there are usually some restrictions and / or certain starting values that are “better” than others in the sense that the algorithm will converge faster, find a more “optimal” solution etc.
- In our case, we can assign our starting values as follows:

$$\beta^{[0]} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$
Step 2: IRLS algorithm

2. Using the re-weighting equation (described next slide), update the $\beta^{[t]}$ vector.

- At step 2, we will update (= produce a new value of the vector) using the following equation (then do this again and again until we stop!):

$$
\beta^{[t+1]} = \beta^{[t]} + [x^T W x]^{-1} x^T (y - \gamma^{-1}(x \beta^{[t]}))
$$

$$
x = \begin{bmatrix}
1 & x_{1,a} & x_{1,d} \\
1 & x_{2,a} & x_{2,d} \\
\vdots & \vdots & \vdots \\
1 & x_{n,a} & x_{n,d}
\end{bmatrix} \quad \beta^{[t]} = \begin{bmatrix}
\beta_{\mu}^{[t]} \\
\beta_{a}^{[t]} \\
\beta_{d}^{[t]}
\end{bmatrix}
$$

$$
y = \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix} \quad W_{ii} = \gamma^{-1}(\beta_{\mu}^{[t]} + x_{i,a} \beta_{a}^{[t]} + x_{i,d} \beta_{d}^{[t]})(1 - \gamma^{-1}(\beta_{\mu}^{[t]} + x_{i,a} \beta_{a}^{[t]} + x_{i,d} \beta_{d}^{[t]}))
$$

$$
W_{ii} = \frac{e^{x_i \beta^{[t]}}}{1 + e^{x_i \beta^{[t]}}}
$$

$$
(W_{ij} = 0 \text{ for } i \neq j)
$$
Step 3: IRLS algorithm

3. At each step \( t > 0 \) check if \( \beta^{[t+1]} \approx \beta^{[t]} \) (i.e. if these are approximately equal) using an appropriate function. If the value is below a defined threshold, stop. If not, repeat steps 2, 3.

- At step 3, we “check” to see if we should stop the algorithm and, if we decide not to stop, we go back to step 2.

- If we decide to stop, we will assume the final values of the vector are the MLE (it may not be exactly the true MLE, but we will assume that it is close if we do not stop the algorithm to early!), e.g. \( \beta^{[t+1]} \approx \beta^{[t]} \)

- There are many stopping rules, using change in Deviance is one way to construct a rule (note the issue with \( \ln(0) \)!!:

\[
\triangle D = |D^{[t+1]} - D^{[t]}| \quad \triangle D < 10^{-6}
\]

\[
D = 2 \sum_{i=1}^{n} y_i \ln \left( \frac{y_i}{\gamma^{-1}(\beta^{[t+1]} + x_{i,a} \beta^a_t + x_{i,d} \beta^d_t)} \right) + (1 - y_i) \ln \left( \frac{1 - y_i}{1 - \gamma^{-1}(\beta^{[t+1]} + x_{i,a} \beta^a_t + x_{i,d} \beta^d_t)} \right)
\]

\[
D = 2 \sum_{i=1}^{n} y_i \ln \left( \frac{y_i}{e^{\beta^{[t+1]} + x_{i,a} \beta^a_t + x_{i,d} \beta^d_t}} \right) + (1 - y_i) \ln \left( \frac{1 - y_i}{1 - e^{\beta^{[t+1]} + x_{i,a} \beta^a_t + x_{i,d} \beta^d_t}} \right)
\]
Logistic hypothesis testing I

- Recall that our null and alternative hypotheses are:
  \[ H_0 : \beta_a = 0 \cap \beta_d = 0 \]
  \[ H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]
- We will use the LRT for the null (0) and alternative (1):
  \[ LRT = -2ln\Lambda = -2ln \frac{L(\hat{\theta}_0|y)}{L(\hat{\theta}_1|y)} \quad LRT = -2ln\Lambda = 2l(\hat{\theta}_1|y) - 2l(\hat{\theta}_0|y) \]
- For our case, we need the following:
  \[ l(\hat{\theta}_1|y) = l(\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d|y) \]
  \[ l(\hat{\theta}_0|y) = l(\hat{\beta}_\mu, 0, 0|y) \]
Logistic hypothesis testing II

- For the alternative, we use our MLE estimates of our logistic regression parameters we get from our IRLS algorithm and plug these into the log-like equation

\[
l(\hat{\theta}_1 | y) = \sum_{i=1}^{n} \left[ y_i \ln(\gamma^{-1}(\beta_\mu + x_{i,a} \hat{\beta}_a + x_{i,d} \hat{\beta}_d)) + (1 - y_i) \ln(1 - \gamma^{-1}(\beta_\mu + x_{i,a} \hat{\beta}_a + x_{i,d} \hat{\beta}_d)) \right]
\]

\[
\gamma^{-1}(\beta_\mu + x_{i,a} \hat{\beta}_a + x_{i,d} \hat{\beta}_d) = \frac{e^{\beta_\mu + x_{i,a} \hat{\beta}_a + x_{i,d} \hat{\beta}_d}}{1 + e^{\beta_\mu + x_{i,a} \hat{\beta}_a + x_{i,d} \hat{\beta}_d}}
\]

- For the null, we plug in the following parameter estimates into this same equation

\[
l(\hat{\theta}_0 | y) = \sum_{i=1}^{n} \left[ y_i \ln(\gamma^{-1}(\beta_{\mu,0} + x_{i,a} \ast 0 + x_{i,d} \ast 0)) + (1 - y_i) \ln(1 - \gamma^{-1}(\beta_{\mu,0} + x_{i,a} \ast 0 + x_{i,d} \ast 0)) \right]
\]

- where we use the same IRLS algorithm to provide estimates of by running the algorithm EXACTLY the same with \( \hat{\beta}_{\mu,0} \) EXCEPT we set \( \hat{\beta}_a = 0, \hat{\beta}_d = 0 \) and we do not update these!
Logistic hypothesis testing III

- To calculate our p-value, we need to know the distribution of our LRT statistic under the null hypothesis.
- There is no simple form for this distribution for any given n (contrast with F-statistics!!) but we know that as n goes to infinite, we know the distribution is i.e. \( n \to \infty \):

\[
LRT = -2\ln \Lambda = 2l(\hat{\theta}_1|y) - 2l(\hat{\theta}_0|y)
\]

\[
LRT \to \chi^2_{df}
\]

- What’s more, it is a reasonably good assumption that under our (not all!!) null, this LRT is (approximately!) a chi-square distribution with 2 degrees of freedom (d.f.) assuming n is not too small!
Logistic Regression p-value

- To calculate our p-value, we need to know the distribution of our LRT statistic under the null hypothesis.
- There is no simple form for this distribution for any given n (contrast with F-statistics!!) but we know that as n goes to infinite, we know the distribution is i.e. \( (n \rightarrow \infty) \):
  \[
  LRT = -2 \ln \Lambda = 2l(\hat{\theta}_1 | y) - 2l(\hat{\theta}_0 | y)
  \]
  \[
  LRT \rightarrow \chi^2_{df}
  \]
Modeling logistic covariates

• Therefore, if we have a factor that is correlated with our phenotype and we do not handle it in some manner in our analysis, we risk producing false positives AND/OR reduce the power of our tests!

• The good news is that, assuming we have measured the factor (i.e. it is part of our GWAS dataset) then we can incorporate the factor in our model as a covariate:

\[ Y = \gamma^{-1}(\beta_\mu + X_a\beta_a + X_d\beta_d + X_z\beta_z) \]

• The effect of this is that we will estimate the covariate model parameter and this will account for the correlation of the factor with phenotype (such that we can test for our marker correlation without false positives / lower power!)

\[ \varepsilon \sim N(0, \sigma^2) \]
Modeling logistic covariates II

• For our a logistic regression, our LRT (logistic) we have the same equations:

\[ LRT = -2ln \Lambda = 2l(\hat{\theta}_1|y) - 2l(\hat{\theta}_0|y) \]

\[
l(\hat{\theta}_1|y) = \sum_{i=1}^{n} \left[ y_iln(\gamma^{-1}(\hat{\beta}_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d + x_{i,z}\hat{\beta}_z)) + (1 - y_i)ln(1 - \gamma^{-1}(\hat{\beta}_\mu + x_{i,a}\hat{\beta}_a + x_{i,d}\hat{\beta}_d + x_{i,z}\hat{\beta}_z)) \right] \]

\[
l(\hat{\theta}_0|y) = \sum_{i=1}^{n} \left[ y_iln(\gamma^{-1}(\hat{\beta}_\mu + x_{i,z}\hat{\beta}_z)) + (1 - y_i)ln(1 - \gamma^{-1}(\hat{\beta}_\mu + x_{i,z}\hat{\beta}_z)) \right] \]

• Using the following estimates for the null hypothesis and the alternative making use of the IRLS algorithm (just add an additional parameter!):

\[ \hat{\theta}_0 = \{ \hat{\beta}_\mu, \hat{\beta}_a = 0, \hat{\beta}_d = 0, \hat{\beta}_z \} \]

\[ \hat{\theta}_1 = \{ \hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d, \hat{\beta}_z \} \]

• Under the null hypothesis, the LRT is still distributed as a Chi-square with 2 degree of freedom (why?):

\[ LRT \rightarrow \chi^2_{df=2} \]
Logistic covariates: summary

- For covariates: say you need to include a single “Z” (note: same structure for more than one)
we start with the same hypotheses:

\[ H_0 : \beta_a = 0 \cap \beta_d = 0 \quad H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]

- We need the logistic model for this case

\[ Y_i = \gamma^{-1}(X\beta) + \epsilon_i \quad Y_i = \frac{e^{\beta_\mu + x_i,a \beta_a + x_i,d \beta_d + x_i,z \beta_z}}{1 + e^{\beta_\mu + x_i,a \beta_a + x_i,d \beta_d + x_i,z \beta_z}} + \epsilon_i \]

- And the associated likelihood equation

\[ l(\beta) = \sum_{i=1}^{n} [y_i \ln(\gamma^{-1}(\beta_\mu + x_i,a \beta_a + x_i,d \beta_d + x_i,z \beta_z)) + (1 - y_i) \ln(1 - \gamma^{-1}(\beta_\mu + x_i,a \beta_a + x_i,d \beta_d + x_i,z \beta_z))] \]

- Where we need to substitute the \( MLE(\hat{\beta}) = MLE(\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d, \hat{\beta}_z) \) for the following two cases:

\[ l(\hat{\theta}_1|y) = \sum_{i=1}^{n} [y_i \ln(\gamma^{-1}(\hat{\beta}_\mu + x_i,a \hat{\beta}_a + x_i,d \hat{\beta}_d + x_i,z \hat{\beta}_z)) + (1 - y_i) \ln(1 - \gamma^{-1}(\hat{\beta}_\mu + x_i,a \hat{\beta}_a + x_i,d \hat{\beta}_d + x_i,z \hat{\beta}_z))] \]

\[ l(\hat{\theta}_0|y) = \sum_{i=1}^{n} [y_i \ln(\gamma^{-1}(\hat{\beta}_\mu + x_i,z \hat{\beta}_z)) + (1 - y_i) \ln(1 - \gamma^{-1}(\hat{\beta}_\mu + x_i,z \hat{\beta}_z))] \]

- So use the same IRLS algorithm with the appropriate equation and \( X \) matrix with new columns
(run the algorithm twice as before!)

- Substitute the MLEs, calculate the \( LRT = -2ln\Lambda = 2l(\hat{\theta}_1|y) - 2l(\hat{\theta}_0|y) \)

- And use a \( LRT \to \chi^2_{df} \) to calculate the p-val!
We have introduced linear and logistic regression models for GWAS analysis because these are the most versatile framework for performing a GWAS (there are many less versatile alternatives!)

These two models can handle our genetic coding (in fact any genetic coding) where we have discrete categories (although they can also handle $X$ that can take on a continuous set of values!)

They can also handle (the sampling distribution) of phenotypes that have normal (linear) and Bernoulli error (logistic)

How about phenotypes with different error (sampling) distributions? Linear and logistic regression models are members of a broader class called Generalized Linear Models (GLMs), where other models in this class can handle additional phenotypes (error distributions)
Introduction to Generalized Linear Models (GLMs) II

• To introduce GLMs, we will introduce the overall structure first, and second describe how linear and logistic models fit into this framework.

• There is some variation in presenting the properties of a GLM, but we will present them using three (models that have these properties are considered GLMs):
  • The probability distribution of the response variable \( Y \) conditional on the independent variable \( X \) is in the exponential family of distributions:
    \[
    \Pr(Y|X) \sim \text{expfamily}
    \]
  • A link function relating the independent variables and parameters to the expected value of the response variable (where we often use the inverse!!)
    \[
    \gamma : E(Y|X) \rightarrow X\beta,
    \gamma(E(Y|X)) = X\beta
    E(Y|X) = \gamma^{-1}(X\beta)
    \]
  • The error random variable \( \epsilon \) has a variance which is a function of ONLY \( X\beta \)
Exponential family I

• The exponential family is includes a broad set of probability distributions that can be expressed in the following `natural’ form:

\[ Pr(Y) \sim e^{\frac{Y \theta - b(\theta)}{\phi} + c(Y, \phi)} \]

• As an example, for the normal distribution, we have the following:

\[ \theta = \mu, \phi = \sigma^2, b(\theta) = \frac{\theta^2}{2}, c(Y, \phi) = -\frac{1}{2} \left( \frac{Y^2}{\phi} + \log(2\pi\phi) \right) \]

• Note that many continuous and discrete distributions are in this family (normal, binomial, poisson, lognormal, multinomial, several categorical distributions, exponential, gamma distribution, beta distribution, chi-square) but not all (examples that are not!!) and since we can model response variables with these distributions, we can model phenotypes with these distributions in a GWAS using a GLM (!!)

• Note that the normal distribution is in this family (linear) as is Bernoulli or more accurately Binomial (logistic)
Exponential family II

• Instead of the `natural’ form, the exponential family is often expressed in the following form:

\[ Pr(Y) \sim h(Y)s(\theta)e^{\sum_{i=1}^{k} w(\theta)t_i(Y)} \]

• To convert from one to the other, make the following substitutions:

\[ k = 1, h(Y) = e^{c(Y,\phi)}, s(\theta) = e^{-\frac{b(\theta)}{\phi}}, w(\theta) = \frac{\theta}{\phi}, t(Y) = Y \]

• Note that the dispersion parameter is now no longer a direct part of this formulation

• Which is used depends on the application (i.e., for glm’s the `natural’ form has an easier to use form + the dispersion parameter is useful for model fitting, while the form on this slide provides advantages for other types of applications)
GLM link function

- A “link” function is just a function (!!) that acts on the expected value of $Y$ given $X$:

$$Y = f(X) \quad f^{-1}(Y) = X$$

- This function is defined in such a way such that it has a useful form for a GLM although there are some general restrictions on the form of this function, the most important is that they need to be monotonic such that we can define an inverse:

$$Y = f(X) \quad f^{-1}(Y) = X$$

- For the logistic regression, we have selected the following link function, which is a logit function (a “canonical link”) where the inverse is the logistic function (but note that others are also used for binomial response variables):

$$\gamma(E(Y|X)) = \ln\left(\frac{e^{X\beta}}{1+e^{X\beta}}\right) \quad E(Y|X) = \gamma^{-1}(X\beta) = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

- What is the link function for a normal distribution?
GLM error function

- The variance of the error term in a GLM must be function of ONLY the independent variable and beta parameter vector:

\[ V\text{ar}(\epsilon) = f(X\beta) \]

- This is the case for a linear regression (note the variance of the error is constant!!):

\[ \epsilon \sim N(0, \sigma_\epsilon^2) \]

\[ V\text{ar}(\epsilon) = f(X\beta) = \sigma_\epsilon^2 \]

- As an example, this is the case for the logistic regression (note the error changes depending on the value of X!!):

\[ V\text{ar}(\epsilon) = \gamma^{-1}(X\beta)(1 - \gamma^{-1}(X\beta)) \]

\[ V\text{ar}(\epsilon_i) = \gamma^{-1}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)(1 - \gamma^{-1}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)) \]
Inference with GLMs

• We perform inference in a GLM framework using the same approach, i.e. MLE of the beta parameters using an IRLS algorithm (just substitute the appropriate link function in the equations, etc.)

• We can also perform a hypothesis test using a LRT (where the sampling distribution as the sample size goes to infinite is chi-square)

• In short, what you have learned can be applied for most types of regression modeling you will likely need to apply (!!)
That’s it for today

- Next OPTIONAL lectures: Bayesian Statistics (!!)