**Review: 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. $(n \to \infty)$:

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

$$LRT \to \chi^2_{df}$$

![Graph showing chi-square distribution for various degrees of freedom](image)
Review: Logistic hypothesis testing

- Recall that our null and alternative hypotheses are:

\[ \begin{align*}
  H_0 & : \beta_a = 0 \cap \beta_d = 0 \\
  H_A & : \beta_a \neq 0 \cup \beta_d \neq 0
\end{align*} \]

- We will use the LRT for the null (0) and alternative (1):

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

- For our case, we need the following:

\[ \begin{align*}
  l(\hat{\theta}_1 | y) &= l(\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d | y) \\
  l(\hat{\theta}_0 | y) &= l(\hat{\beta}_\mu, 0, 0 | y)
\]
Review: 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

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

\[
\gamma^{-1}(\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d) = \frac{e^{\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d}}{1 + e^{\beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d}}
\]

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

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

• 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!
Review: 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!
Review: logistic covariates 1

• 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!)
Introduction to Generalized Linear Models (GLMs) I

• 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)
• 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?):

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

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

• 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 $X$:

- 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):

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

- What is the link function for a normal distribution?
The variance of the error term in a GLM must be function of ONLY the independent variable and beta parameter vector:

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

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

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

\[ \text{Var}(\epsilon) = f(X\beta) = \sigma^2_\epsilon \]

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

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

\[ \text{Var}(\epsilon_i) = \gamma^{-1}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)(1 - \gamma^{-1}(\beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d)) \]
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 (!!)