

BTRY 4830/6830: Quantitative Genomics and Genetics Fall 2014

Homework 5 (version 2 - posted October 29, minor edits)

Assigned October 24; Due 11:59PM October 30

Problem 1 (Easy)

- Explain the two major differences between the ‘error’ ($= \epsilon$) term in a logistic regression as compared to a linear regression.
- Recall that a response variable $Y|X$ in a generalized linear model (GLM) has to have a probability distribution in the exponential family, i.e. $Y|X \sim \text{expfamily}(\theta)$. We defined a random variable as having a probability distribution in the exponential family if the pdf can be expressed in terms of the following equation:

$$Pr(Y) \sim e^{\frac{Y\theta - b(\theta)}{\phi} + c(Y, \phi)} \quad (1)$$

Show that with the following substitutions, you can produce a normal distribution (show your steps!):

$$\theta = \mu, \phi = \sigma^2, b(\theta) = \frac{\theta^2}{2}, c(Y, \phi) = -\frac{1}{2} \left(\frac{Y^2}{\phi} + \ln(2\pi\phi) \right) \quad (2)$$

Hint: The following relationships may be useful: $e^a e^b = e^{a+b}$, $\ln(a/b) = \ln(a) - \ln(b)$, $\ln(a^b) = b * \ln(a)$ $e^{\ln(a)} = a$.

Problem 2 (Medium)

For this problem, you will need to write your own code to simulate data for a population of $n = 100$ individuals considering a SINGLE genotype and a phenotype assuming a logistic regression model with a response variable with two states ($Y = 0$ or 1). You will be asked to do this for a case where the ACTUAL case is that the null hypothesis is true and, a case separate case, where the null hypothesis is false. You will then need to analyze these two simulated datasets assuming that you do NOT know what the actual (=correct) answer is in either case. For many parts of this question, the answer will be your R code presented in a text file that is easy for us to run to produce the requested output (NOTE THAT FOR FULL CREDIT = easy to run code in a txt file, name your files appropriately, send a zip file, and do not copy someone else’s code!! etc.).

- Write R code to create a 100×1 matrix of genotypes (i.e. just one locus!) for the $n = 100$ individuals, where there should be two alleles (you must represent these with two of ‘a’,

‘c’, ‘t’, ‘g’ OR as ‘A1’ and ‘A2’) where all three genotype states (two homozygotes and a heterozygote) should be present and the minor allele frequency (MAF) should not be < 0.2 . Write R code to convert your genotype into the appropriate x_a values and the appropriate x_d values. (note that you may use modified code you have already constructed in a previous homework as your answer!)

- b. Write R code to simulate a 100 x 1 vector of y values (either ‘0’ or ‘1’) using the genotypes of part [a] of the 100 individuals under the logistic regression model. Use your code to simulate TWO different 100 x 1 phenotypes (again use the same genotypes from part [a] in each case): for the first, use the parameter values $\beta_\mu = 1.0, \beta_a = 0, \beta_d = 0$ and for the second, use the parameter values $\beta_\mu = 1.0, \beta_a = 1.0, \beta_d = -1.0$.

NOTE: the combined results of parts [a & b] will result in TWO datasets of $n = 100$ where EACH dataset has one marker and one phenotype (and EACH dataset will therefore have 100x1 values for x_a, x_d , and y). While you will consider these two separate datasets, they will actually have the same genotypes in both cases. The first dataset (DATA1) will be a case where the null hypothesis is true and the second datasets (DATA2) will be a case where the null hypothesis is false.

- c. For EACH data set you simulated [DATA1 & DATA2] produce two plots (four plots in total!) using the package ‘ggplot2’ with the option ‘position = position_jitter(w=0.1, h=0.1)’: for the first, plot y versus x_a and for the second, plot y versus x_d .
- d. For each data set [DATA1 & DATA2], calculate the $MLE(\hat{\beta})$ using the IRLS algorithm (make sure you provide the code for you algorithm!). How close are your estimates to the true values you simulated, i.e. present $\hat{\beta} - \beta = [\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d] - [\beta_\mu, \beta_a, \beta_d]$ for each of your β parameters? Explain why these are the estimates of the model parameters under the alternative hypothesis θ_1 ?
- e. For each data set [DATA1 & DATA2] use your IRLS algorithm to calculate the $MLE(\hat{\beta})$ under the null hypothesis. How close are your estimates to the true values you simulated, i.e. present $\hat{\beta}_\mu - \beta_\mu$? Explain why these are the estimates of the model parameters under the null hypothesis θ_0 ?
- f. Why is it likely that the parameter estimates you obtained in part [e] are not exactly the true $MLE(\hat{\beta})$ but are probably close?
- g. For each data set [DATA1 & DATA2], calculate the $LRT = -2\ln\Lambda$ statistic. Which data set produced a greater value of this statistic?
- h. For the LRT ’s you calculated in part [g] calculate the p-value under the genetic null hypothesis ($H_0 : \beta_a = 0 \cap \beta_d = 0$) versus the alternative ($H_0 : \beta_a \neq 0 \cup \beta_d \neq 0$) assuming that when the null hypothesis is true, the LRT is distributed as a $\chi_{df=2}^2$. For which data sets did you reject the null hypothesis assuming an $\alpha = 0.05$ (provide the p-values you obtained as part of your answer)?
- i. What are two reasons why the p-value you obtained in part [h] are not likely to be the exact p-values (but are probably close)? (hint: one of the reasons is the answer to part [f]!)
- j. For each data set [DATA1 & DATA2], why do the expected values of the phenotype, given each possible genotype state, as calculated using your $MLE(\hat{\beta})$ estimates in part [d] make sense?

What is the interpretation of these expected values of the phenotype and which genotype in which data set would be of greatest concern to an individual given this interpretation?

Problem 3 (Difficult)

- a. Starting for the form of the pdf of distributions in the exponential family:

$$Pr(Y) \sim e^{\frac{Y\theta - b(\theta)}{\phi} + c(Y, \phi)} \quad (3)$$

What are the values of $\theta, \phi, b(\theta), c(Y, \phi)$ for the binomial distribution and perform the substitutions and show the steps needed to produce the standard form of a binomial pdf.

- b. Technically, equation (3) is the ‘natural form’ of the equation describing exponential families, which includes the additional ‘dispersion’ parameter ϕ . You will often see the exponential family written using another formula:

$$Pr(Y) \sim h(Y)s(\theta)e^{\sum_{i=1}^k w_i(\theta)t_i(Y)} \quad (4)$$

What are the values of $k, h(Y), s(\theta), w(\theta), t(Y)$ needed to express equation (4) in the form of equation (3), also perform the substitutions and show the steps needed.