Applied Biostatistics

https://moodle.epfl.ch/course/view.php?id=15590

- Statistical modeling overview
- Generalized linear modeling
- Binary data and logistic regression
- Count data and Poisson regression
- Comparing models

Modeling overview

- Want to capture important features of the *relationship between* a (set of) *variable(s)* and one or more *response(s)*
- Many models are of the form

$$g(Y) = f(\mathbf{x}) + \text{error}$$

- *Differences* in the form of g, f and distributional assumptions about the error term

Examples of models

- Linear : $Y = \beta_0 + \beta_1 x + \epsilon$
- Linear : $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
- (Intrinsically) Nonlinear : $Y = \alpha x_1^{\beta} x_2^{\gamma} x_3^{\delta} + \epsilon$
- Generalized Linear Model (e.g. Binomial) :

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x + \beta_2 x_2$$

- Proportional Hazards (in Survival Analysis) :

$$h(t) = h_0(t) \exp(\beta x)$$

Linear modeling

- A simple linear model : $E(Y) = \beta_0 + \beta_1 x$
- Gaussian measurement model : $Y = \beta_0 + \beta_1 x + \epsilon, \epsilon\ N(0, \sigma^2)$
- More generally : $Y = X\beta + \epsilon$, where Y is $n \times 1$, X is $n \times p$, β is $p \times 1$, ϵ is $n \times 1$, often assumed $N(0, \sigma^2 I_{n \times n})$

Analysis of designed experiments

- An important use of linear models – we have already done this using anova
- Define a (design) matrix X so that for response variable Y :

$$E(Y) = X\beta,$$

  where β is a vector of *parameters* (or contrasts)
- Many ways to define design matrix/contrasts

Model fitting and checking

- For the standard (*fixed effects*) linear model, estimation is usually by *least squares*
- Can be more complicated with *random effects* or when x-variables are subject to measurement error as well
- Checking model : examination of *residuals*
  - Normality
  - Time effects
  - Nonconstant variance
  - Curvature
- Detection of *influential observations*

Linear regression model (again)

- Linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- Another way to write this :

$$Y \sim N(\mu, \sigma^2), \quad \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- Suitable for a *continuous* response
- **NOT** suitable for a *binary* response
- **NOT** suitable for a *count* data

Modified model

- Instead of modeling the response directly, could instead model some *function* of the response
- i.e., Instead of modeling the expected response *directly* as a linear model, model a *suitable transformation*
- For binary data, it is convenient to use the *logit* function
- For count data, this is often taken to be the *log* transformation

Modified model for binary data

- Instead of modeling the 0/1 response directly, could instead model the *probability* of '1'
- Problems :
  - could lead to fitted values outside of $[0, 1]$
  - normality assumption on errors is wrong
- Instead of modeling the expected response *directly* as a linear function of the predictors, model a *suitable transformation*
- For binary data, this is generally taken to be the *logit* (or *logistic*) transformation

Logit transformation

- $\text{logit}(p) = \log \dfrac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$
- Therefore,

$$p(x_1, \ldots x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$$

- The parameter β_k is such that $\exp(\beta_k)$ is the *odds* that the response takes value 1 when x_k increases by one, when the remaining variables are constant
- *i.e.* β_k is a *log-odds*
- Estimate parameters by *maximum likelihood* rather than least squares

Generalized linear model

- In a standard linear model, the *response variable* is modeled as a *normally distributed*
- However, if the response variable is *dichotomous* or a *count*, it does not make sense to model the outcome as normal
- Generalized linear models (GLMs) are an extension of linear models to model non-normal response variables
- A GLM consists of three components :
  - A *random component*, specifying the conditional distribution of the response variable, Y_i, given the values of the explanatory variables in the model
  - A *linear predictor*
  - A smooth and invertible linearizing *link function*
- We consider *logistic regression* for a count response
- We can consider *Poisson regression* for a count response

Generalized linear models : some theory

- Allows unified treatment of statistical methods for several important classes of models
- Response Y assumed to have *exponential family distribution* :

$$f(y) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

- For a standard linear model

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon, \text{ with } \epsilon \sim N(0, \sigma^2)$$

- The *expected response* is $E[Y \mid x] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- Let η denote the *linear predictor* $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- For a standard linear model, $E[Y \mid x] = \eta$
- In a *generalized linear model*, there is a *link function* g between η and the expected response :

$$g(E[Y \mid x]) = \eta$$

- For a standard linear model, $g(y) = y$ (*identity link*)

Link function

- When the response variable is binary (with values coded as 0 or 1), then $E[Y \mid x] = P(Y = 1 \mid x)$
- A convenient function in this case is
$$E[Y \mid x] = P(Y = 1 \mid x) = \frac{e^{\eta}}{1 + e^{\eta}}$$
- The corresponding link function (inverse of this function) is called the *logit*
- $\text{logit}(x) = \log \frac{x}{1 - x}$
- Regression using this model is called *logistic regression*

Link function : examples

	Family Name				
Link	binomial	Gamma	gaussian	inverse.gaussian	poisson
logit	D				
probit	•				
cloglog	•				
identity		•	D		•
inverse		D			
log		•			D
1/mu^2				D	
sqrt					•

Analogous to linear regression

- The logit function g has many of the desirable properties of a linear regression model :
  - Mathematically convenient and flexible
  - Can meaningfully interpret parameters
  - Linear in the parameters
- A difference : Error distribution is binomial (**not** normal)

Fitting the model

- For linear regression, typically use *least squares*
- For dichotomous data or count data, the 'nice' statistical properties of least squares estimators no longer hold
- The general estimation method that leads to least squares (for normally distributed errors) is *maximum likelihood*
- Write out the likelihood, take the derivative, set equal to zero and solve
- Estimating equations typically nonlinear functions of the regression parameters so must be solved numerically (IRLS)

Maximum likelihood estimation

- Likelihood : $f(x_i) = p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i}$
- Assuming independent observations, the likelihood $l(\beta) = \prod_{i=1}^{n} f(x_i)$
- log likelihood $L(\beta) = \log[l(\beta)] = \sum_{i=1}^{n}(y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i)))$
- To find β that maximize the log likelihood, differentiate wrt each β_i and set the derivative equal to 0
- In linear regression these equations are easily solved
- In logistic regression, these are nonlinear in β and are solved iteratively

PAUSE

DNA sequencing

- (Automated) Sanger sequencing
  - 'first-generation' technology
  - F. Sanger, 1977
- Process :
  - bacterial cloning or PCR
  - template purification
  - labelling of DNA fragments using the chain termination method with energy transfer, dye-labelled dideoxynucleotides and a DNA polymerase
  - capillary electrophoresis
  - fluorescence detection
- Data : four-colour plots that reveal the DNA sequence

Next-generation sequencing

- Several newer sequencing technologies
  - 'Next-generation sequencing' (NGS data)
  - 'Ultra high-throughput sequencing' (UHTS data)
- These newer technologies use various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods
- Data : four-colour plots that reveal the DNA sequence
- Major advance : ability to produce a *large amount* of data relatively *cheaply*
- Expands experimental possibilities beyond just determining the order of bases

Applications of NGS

- Sequence assembly (original application)
- Resequencing : The sequencing of part of an individual's genome in order to detect sequence differences between the individual and the standard genome of the species
- Gene expression : RNA-Seq
- SNP discovery and genotyping
- Variant discovery and quantification
- Transcription factor binding sites : ChIP-Seq
- Measuring DNA methylation

NGS data generation

- Sequencing technologies incorporate methods that we can class as
  - template preparation
  - sequencing and imaging
  - data analysis
- Combination of specific protocols distinguishes different technologies
- Major technologies :
  - Illumina HiSeq (older : Solexa)
  - 454 (Roche)
  - Applied Biosciences SOLiD
  - Pacific Biosciences SMRT (single molecule real-time)

Data analysis pipeline

- Data are *counts* of short sequences (called 'reads')
- Quality control of data
- Match to reference sequence, read mapping
- Count/summarize number of reads per feature
- Statistical analysis (depends on the specific application)

Sequence data

- Sequence data are *counts*
- DNA sample $\Longrightarrow$ *population of cDNA fragments*
- Each genomic feature $\Longrightarrow$ species for which the population size is to be estimated
- Sequencing a DNA sample $\Longrightarrow$ random sampling of each of these species
- *Aim :* to estimate the relative abundance of each species in the population

Poisson model

- If we assume :
  - each cDNA fragment has the *same chance* of being selected for sequencing
  - the fragments are selected independently
- Then : the number of read counts for a given genomic feature should follow a *Poisson variation law* across repeated sequence runs of the same cDNA sample
- The Poisson model implies that the *mean equals the variance*
- (This relationship has been validated in an early RNA-Seq study using the same initial source of RNA distributed across multiple lanes of an Illumina GA sequencer)

Single gene model

- DNA sample $\implies$ 'library'
- Contains genes $1, \ldots, g, \ldots$
- For a given gene g in library i, Y_{gi} = number of reads for gene g in library i
- $Y_{gi} \sim Bin(M, p_{gi})$, where p_{gi} is the proportion of the total number of sequences M in library i that are gene g
- M large, p_{gi} small $\implies$ $Y_{gi} \sim Pois(\mu_{gi} = Mp_{gi})$ (approximately)

Technical vs. biological replicates

- For the Poisson model, the *variance* is equal to the *mean*
- With *technical replicates*, this relation holds fairly well
- With *biological replicates*, the variance is typically *larger* than expected using the Poisson model
- There are a few different approaches for accounting for this additional variability (overdispersion)

Link function for count data

- We can model the count data $Y_i \sim Pois(\mu_i),\ i = 1, \ldots, n$
- Want to relate the mean μ_i to one or more *covariates* (for example, treatment/control status)
- A convenient link function in this case is the log :

$$\log \mu_i = \eta = x_i^T \beta$$

- Using a log link ensures that the fitted values of μ_i will remain in the parameter space $[0, \infty)$
- A Poisson model with a log link is sometimes called a *log-linear model*

Variance function for the Poisson model

- The Poisson distributions are a discrete family with probability function indexed by the rate parameter $\mu > 0$:

$$p(y) = \frac{e^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, \ldots$$

- Under the Poisson model : $E[Y_i] = Var(Y_i) = \mu_i$
- General form of the relationship between the variance of the response variable and its mean is : $Var(response) = \phi V(\mu)$, with ϕ a constant scale factor
  - **Normal :** $V(\mu) = 1, \quad \phi = \sigma^2$ (the variance does not depend on the mean)
  - **Binomial :** $V(\mu) = \mu(1 - \mu) \quad \phi = 1$
  - **Poisson :** $V(\mu) = \mu \quad \phi = 1$
- Real data are often *overdispersed*, exhibiting more variation than allowed by the Poisson model

Detecting and handling overdispersion

- When fitting a GLM with binomial or Poisson errors, can often detect overdispersion by *comparing the residual deviance to its degrees of freedom*
- For a well-fitting model, these should be approximately equal
- Overdispersion usually handled with an alternative model :
  - **Quasi-Poisson Model** : Assume $Var(Y_i) = \phi\,\mu_i$ and estimating the *scale parameter* ϕ
  - *Zero-Inflated Poisson Model* : for modeling the case when there are too many '0' values
  - *Negative Binomial Model* : Can arise from a two-stage model :

$$Y_i \sim Pois(\mu_i^*) \qquad \mu_i* \sim \Gamma(\mu_i/\omega, \omega)$$

Then $Y_i \sim NegBin$, with $E[Y_i] = \mu_i$ and $Var(Y_i) = \mu_i + \mu_i^2/\omega$

Differential gene expression for NGS data

- Several BioConductor (R) packages for identifying differential expression from NGS data
- These mostly use the negative binomial model, since the counts are typically over-dispersed compared to the Poisson model
- The `edgeR` package uses an overdispersed Poisson model to account for both biological and technical variability, and uses empirical Bayes methods to moderate the degree of overdispersion across transcripts

Assessing model fit

- In linear regression, an anova table partitions SST, the total sum of squared deviations of observations about their mean, into two parts :
  - SSE, or residual (observed - predicted) sum of squares
  - SSR, or regression sum of squares
- Large SSR suggests the explanatory variable(s) is(are) important
- In linear regression, diagnostics are built around residuals and SSR
- For GLMs, there are a few different kinds of residuals : *Pearson* residuals and *deviance* residuals
- Pearson residual for an observation is obtained by subtracting the mean (predicted value) for that observation and dividing by the (estimated) SD
- Deviance residuals are based on the contribution of each point to the likelihood

Deviance

- In standard linear models, estimate parameters by minimizing residual sum of squares
- (Equivalent to ML for normal model)
- In GLM, estimate parameters by ML
- The *deviance* is (proportional to) $2 \times l$
- (Analogous to SSE)
- Obtaining 'absolute' measure of goodness of fit depends on some assumptions that may not be satisfied in practice
- Usually focus on comparing competing models
- When the models are *nested*, can carry out likelihood ratio test

Comparing models

- In linear regression, consider coefficient significant if (squared) standardized value $\hat{\beta}/SE(\hat{\beta})$ is 'large'
- Can also do this for logistic regression (Wald test), but there are some problems with it
- Preferred approach : likelihood ratio test
- Deviance $D = -2\sum_{i=1}^{n} y_i \log\left(\frac{\hat{p}_i}{y_i}\right) + (1-y_i)\log\left(\frac{1-\hat{p}_i}{1-y_i}\right)$
- To compare models, compute $G = D(\text{submodel}) - D(\text{bigger model})$
- Under the null (*i.e.* the submodel), $G \sim \chi^2$ with df = difference in the number of estimated parameters

Variance inflaction factors

- The meaning of a variance inflation factor is essentially equivalent for linear models and GLMs
- We can use the VIF to look for multicollinearity
- `R` function `vif` from the `car` package
- Also look at correlation matrix for the data matrix X

Summary

- Residuals are certainly less informative for GLMs than for linear regression
- Issues of outliers and influential observations just as relevant for GLMs as for linear regression : look at Cook's distance plot
- Usually a good idea to *start with simple models* and gradually add in complexity