Linear Regression and Generalized Linear Model

Yun (Renee) Zhang, PhD

Email: zhangy@jcvi.org

Monthly Biostatistics Discussion

09-19-2018

Outline

1. Inference for Linear Regression
2. Generalized Linear Model (GLM)
3. Model Fitting and Inference for GLM

Inference for Linear Regression

Significance of Regression Coefficient

- Fitted model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Coefficient estimates can be obtained from statistical software
  - $\hat{\beta}_0$: intercept
  - $\hat{\beta}_1$: slope ("effect size", association, +/- direction)
- Testing hypotheses

$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0$$

- t-test

$$t = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)},$$

  which follows a t-distribution with degrees of freedom (DF) $n - 1$
- For multiple regression, DF = # of observations – # of covariates

Overall Significance of the Regression Model

- Full model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
- Null model: $y = \beta_0$
- Testing hypotheses

$$H_0: \beta_1 = \cdots = \beta_p = 0$$

$$H_1: \beta_j \neq 0 \text{ for at least one } j$$

- F-test

Source	df	SS	MS	F
Regression	p	$SSR = \sum_i (\hat{y}_i - \bar{y})^2$	$MSR = SSR/p$	MSR/MSE
Residual	$n-p-1$	$SSE = \sum_i (y_i - \hat{y}_i)^2$	$MSE = SSE/(n-p-1)$	
Total	$n-1$	$SST = \sum_i (y_i - \bar{y})^2$		

$$F = \frac{\text{explianed variance}}{\text{unexplained variance}},$$

which follows an F-distribution with DFs = p and $n - p - 1$

- For simple linear regression (i.e. only one covariate), t-test and F-test are equivalent, and $F = t^2$

Example

```
## Call:
## lm(formula = dist ~ speed.c, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.9800     2.1750  19.761  < 2e-16 ***
## speed.c       3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Model specification (intercept not shown explicitly)

p-value

t-test

$\hat{\beta}_1$ | $\text{se}(\hat{\beta}_1)$ | $\hat{\beta}_1/\text{se}(\hat{\beta}_1)$

Equal if simple linear regression

F-test

Model Selection

- Full model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
- Reduced model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q,$ where $q < p$
- Testing hypotheses

$$H_0: \beta_{q+1} = \cdots = \beta_p = 0 \text{ (reduced model is true)}$$

$$H_1: \beta_j \neq 0 \text{ for at least one } j \text{ (reduced model is not true)}$$

- F-test

$$F = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/(p-q)}{SSE_{\text{full}}/(n-p-1)}$$

$$= \frac{\text{gain of explained variance by adding additional } p-q \text{ covariates}}{\text{unexplained variance in the full model}},$$

which follows an F-distribution with DFs = $p - q$ and $n - p - 1$

- Useful in stepwise regression to select the best set of covariates

Generalized Linear Model (GLM)

What is a "linear" model?

- In linear models, the term "linear" refers to the linearity of β, the regression coefficients in the model
- Are the following linear models?
  - Model 1:

$$Y = \beta_1 + \beta_2 \left(\frac{x_1}{x_1 + x_2} \right) e^{x_3} + \epsilon$$

  - Model 2:

$$Y = \frac{\beta_1 x}{\beta_2 + x} + \epsilon$$

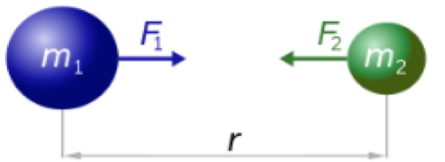
  - Model 3:

$$Y = \beta_1 + \beta_2 e^{-\beta_3 x} + \epsilon$$

  - Model 4:

$$F = G \frac{m_1 \times m_2}{r^{\beta}}$$

Newton's gravity model



$\hat{\beta} = 2$

- Yes; No; No; Yes, $\log F = \log(G m_1 m_2) - \beta \log r$

Linear Model

- A standard linear regression model is

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}_{\text{Systematic component}} + \underbrace{\epsilon}_{\text{Random component}} \qquad \text{where } \epsilon \sim N(0, \sigma^2)$$

- The probability distribution for the standard linear model is

$$Y | x_1, \cdots, x_p \sim N(\mu, \sigma^2)$$

- The essence of the linear model is conditional expectation

$$\mathbb{E}\left(Y \middle| x_1, \cdots, x_p\right) = \beta_0 + \beta_1 x + \cdots + \beta_p x_p$$

- An alternative expression is

$$\mu = \boldsymbol{x}^T \boldsymbol{\beta},$$

where $\boldsymbol{x} = \left(1, x_1, \cdots, x_p\right)^T, \boldsymbol{\beta} = \left(\beta_0, \beta_1, \cdots, \beta_p\right)^T$ are in vector forms

Generalized Linear Model

$$g(\mu) = \boldsymbol{x}^T \boldsymbol{\beta}$$

- Linear predictor: η is a linear combination (thus, "linear") of $\boldsymbol{\beta}$
$$\eta := \boldsymbol{x}^T \boldsymbol{\beta}$$
- Probability distribution: $\mathcal{P}$ is a member of the exponential family
$$Y|\boldsymbol{x} \sim \mathcal{P}(\mu, V(\mu))$$
$$\mathbb{E}(Y|\boldsymbol{x}) = \mu$$
- Link function: $g(\cdot)$ connects η and μ
$$\eta = g(\mu)$$
$$\mu = g^{-1}(\eta)$$
- E.g.
  - Linear regression: $\mathcal{P}$ is a normal distribution, $g(\mu) = \mu$
  - Logistic regression: $\mathcal{P}$ is a Bernoulli distribution, $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$

Basic Concepts of Probability Distribution

- Random variable Y follows probability distribution $\mathcal{P}$

$$Y \sim \mathcal{P}$$

- Discrete distribution / continuous distribution

$$Y = y_1, y_2, \cdots \text{ the values that } Y \text{ can take}$$

- Probability mass function / probability density function

$$p(y) \text{ or } f(y) = \mathbb{P}(Y = y)$$

- Cumulative distribution function

$$F(y) = \mathbb{P}(Y \leq y)$$

- Distribution family: a set of distributions that can be characterized in a certain way, e.g. location-scale family, exponential family

Exponential Family

Mix/separation of data and parameter in this way

- A set of distributions that can be expressed in the following form

$$f(y;\theta,\phi) = \exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\}$$

- Canonical parameter: θ; dispersion parameter: ϕ
- Exponential family distributions includes normal, Bernoulli, binomial, multinomial, Poisson, negative binomial, beta, exponential, Gamma, …
- E.g. $Y \sim \text{Bernoulli}(p)$

$$f(y) = p^y(1-p)^{1-y} = \exp\{y\ln(\frac{p}{1-p}) + \ln(1-p)\}$$

$$\theta = \ln\left(\frac{p}{1-p}\right), b(\theta) = \ln(1+e^\theta)$$

- Therefore, $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$ is a link function for logistic regression

Link Function

- Establishes a relationship between the mean of the distribution (μ) and the linear predictor ($\boldsymbol{x}^T\boldsymbol{\beta}$)
- There is always a well-defined canonical link function which is derived from the canonical parameter of the exponential family distribution
  - $\theta = g(\mu) = \boldsymbol{x}^T\boldsymbol{\beta}$, which eases computation
  - It's always true that $\mu = b'(\theta)$ for exponential family distributions
- Other link functions are available for different purposes
- E.g. logistic regression
  - Logit link (canonical), probit link, complementary log-log link

Common Distributions with Canonical Link Functions

Canonical

Distribution	Support of distribution	Typical uses	Link name	Link function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$
Gamma				
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-2}$
Poisson	integer: $[0, +\infty)$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$
Bernoulli	integer: $[0, 1]$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$
Binomial	integer: $[0, N]$	count of # of "yes" occurrences out of N yes/no occurrences		
Categorical	integer: $[0, K)$	outcome of single K-way occurrence		
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1			
Multinomial	K-vector of integer: $[0, N]$	count of occurrences of different types (1 .. K) out of N total K-way occurrences		

Source: Wikipedia

Regression Models Used in Biological Data

- Genome-wide association study (GWAS) – logistic regression
- Microarray – linear regression
- RNA-seq – negative binomial regression
- Microbiome – beta regression / negative binomial regression

GWAS

Logistic regression: more flexible analysis for GWA studies

- Similar to linear regression, used for binary outcomes instead of continuous outcomes
- Let Y_i be the phenotype for individual i
  - $Y_i = 0$ for controls
  - $Y_i = 1$ for cases
- Let X_i be the genotype of individual i at a particular SNP
  - TT $X_i = 0$
  - GT $X_i = 1$
  - GG $X_i = 2$
- Add extra terms to adjust for potential confounders: e.g. ethnicity, genotyping batch, genotypes at other SNPs

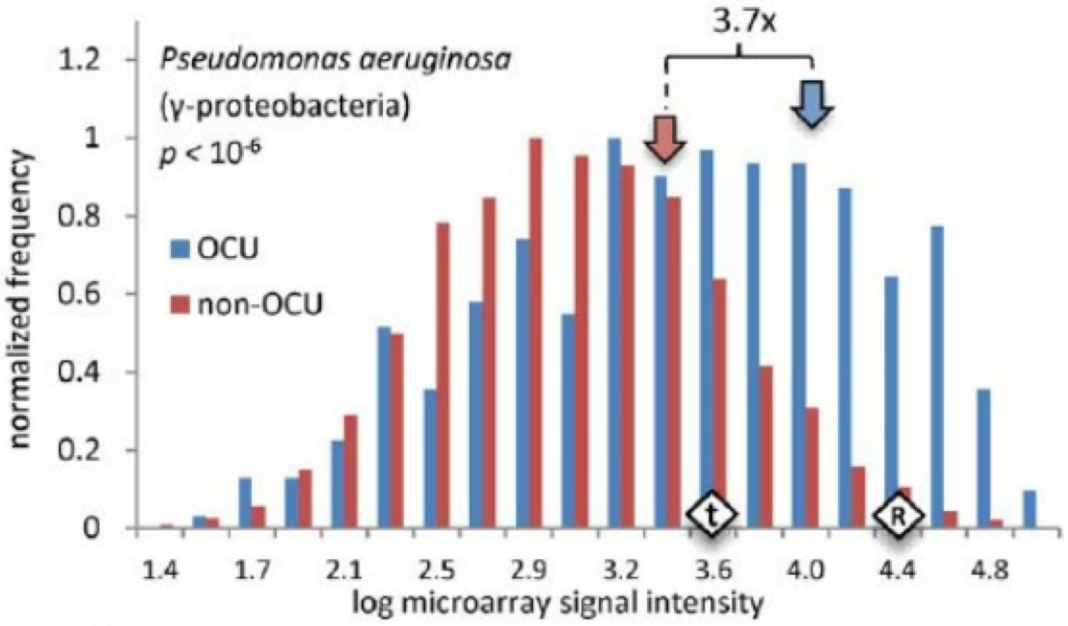
  Let $p_i = E(Y_i \mid X_i, C_i, D_i, \ldots)$

$$\text{logit}(p_i) \sim \beta_0 + \beta_1 X_i + \beta_2 C_i + \beta_3 D_i + \ldots$$

Source: http://bioinformatics.org.au/ws09/presentations/Day3_JStankovich.pdf

Microarray

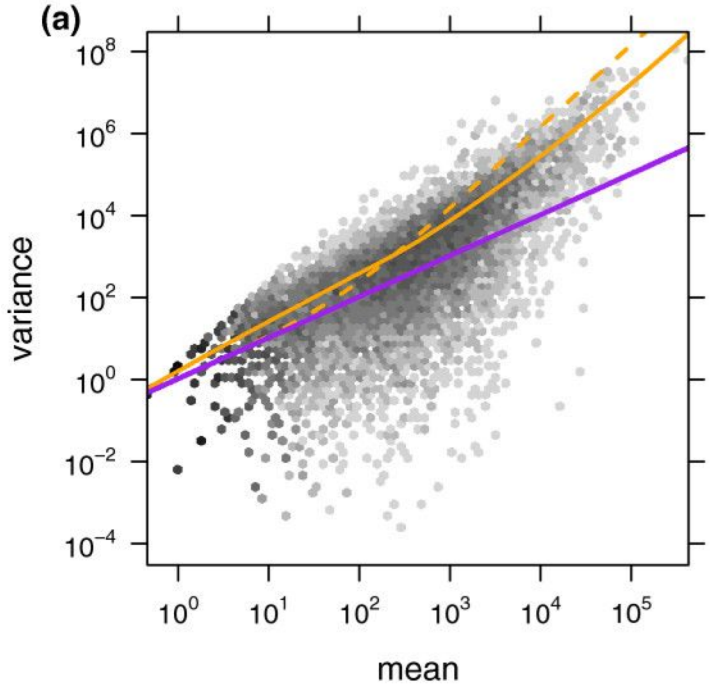
- Microarray intensities are continuous data
- "Bell" shape distribution – normal
- R package: limma



Supek, Fran, et al. "Translational selection is ubiquitous in prokaryotes." PLoS genetics 6.6 (2010): e1001004.

RNA-seq

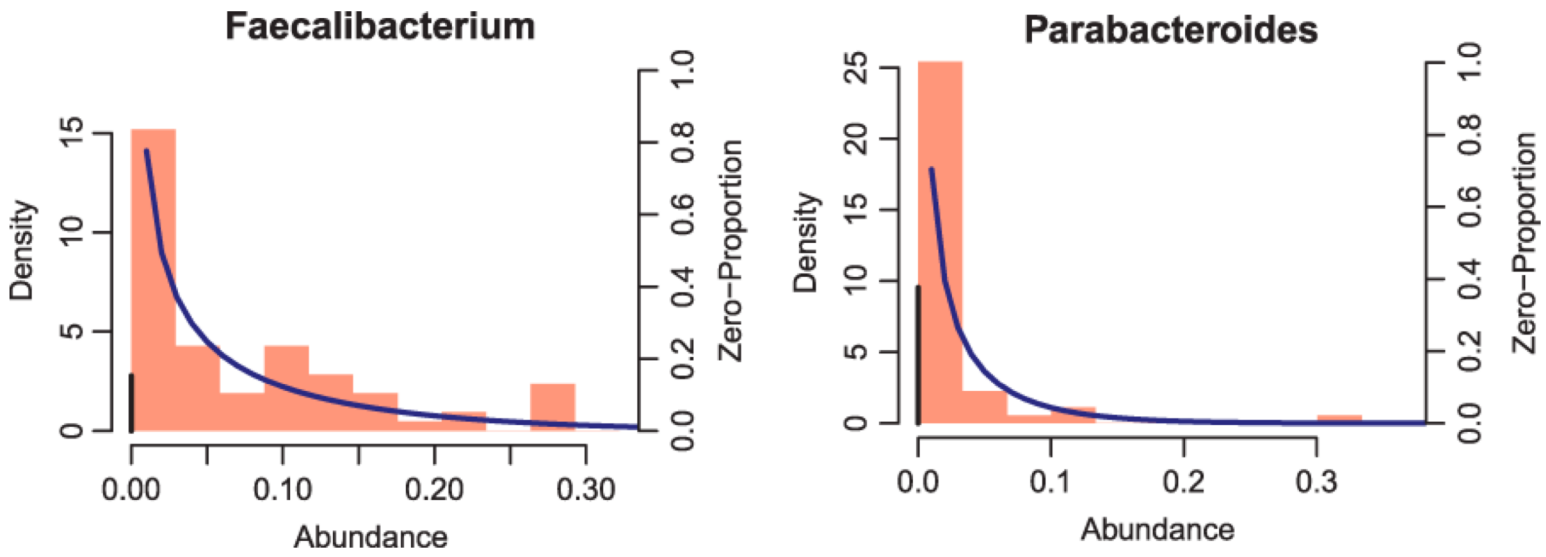
- Sequencing counts are discrete data
- Binomial → Poisson → negative binomial (a.k.a. overdispersed Poisson)
- R package: DEseq2, edgeR



Anders, Simon, and Wolfgang Huber. "Differential expression analysis for sequence count data." Genome biology 11.10 (2010): R106.

Microbiome

- Consider relative abundance of each taxon separately
- Bounded in [0,1); highly skewed – beta distribution

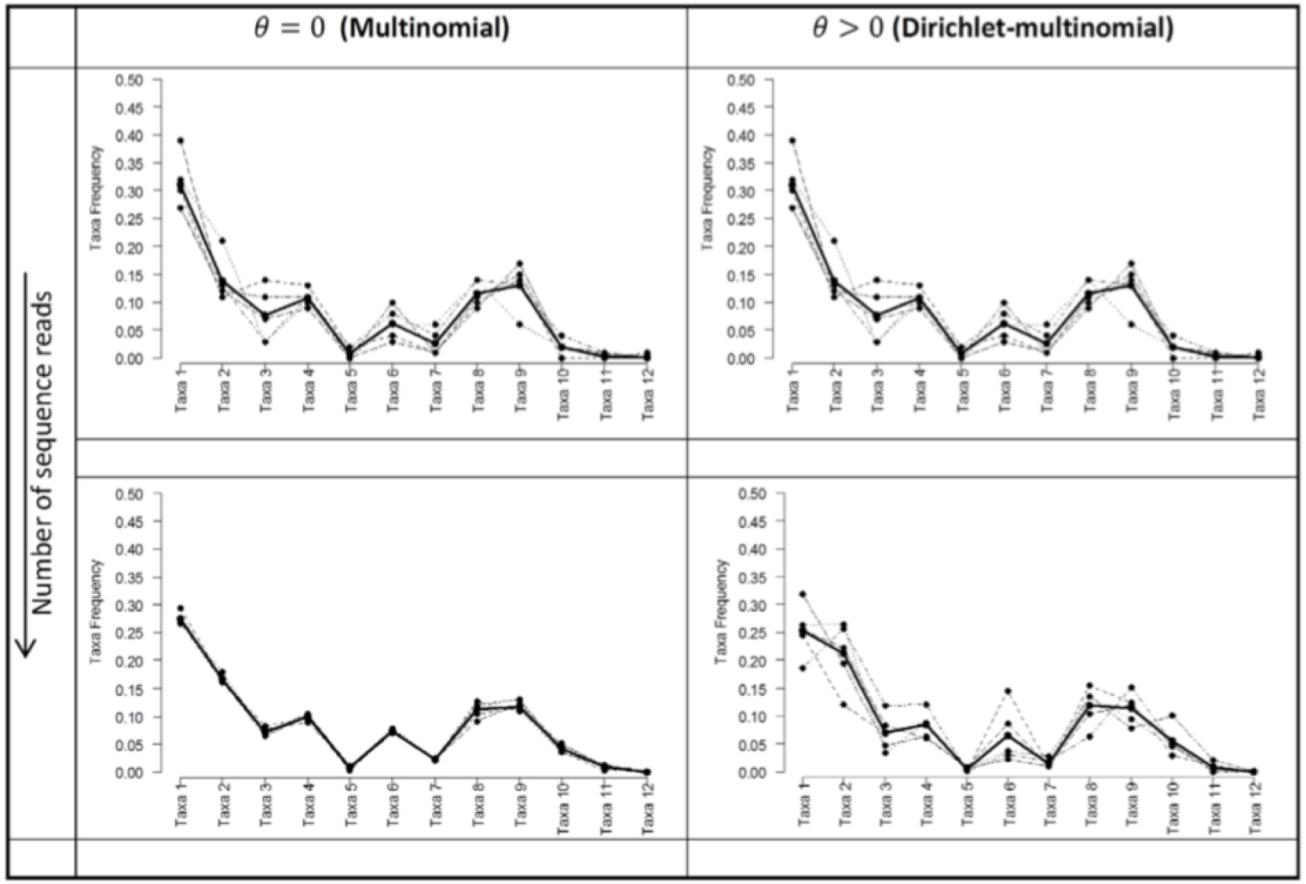


Chen, Eric Z., and Hongzhe Li. "A two-part mixed-effects model for analyzing longitudinal microbiome compositional data." *Bioinformatics*32.17 (2016): 2611-2617.

- If consider absolute abundance, then negative binomial model is a good choice as the nature of sequencing counts

Overdispersion – Microbiome

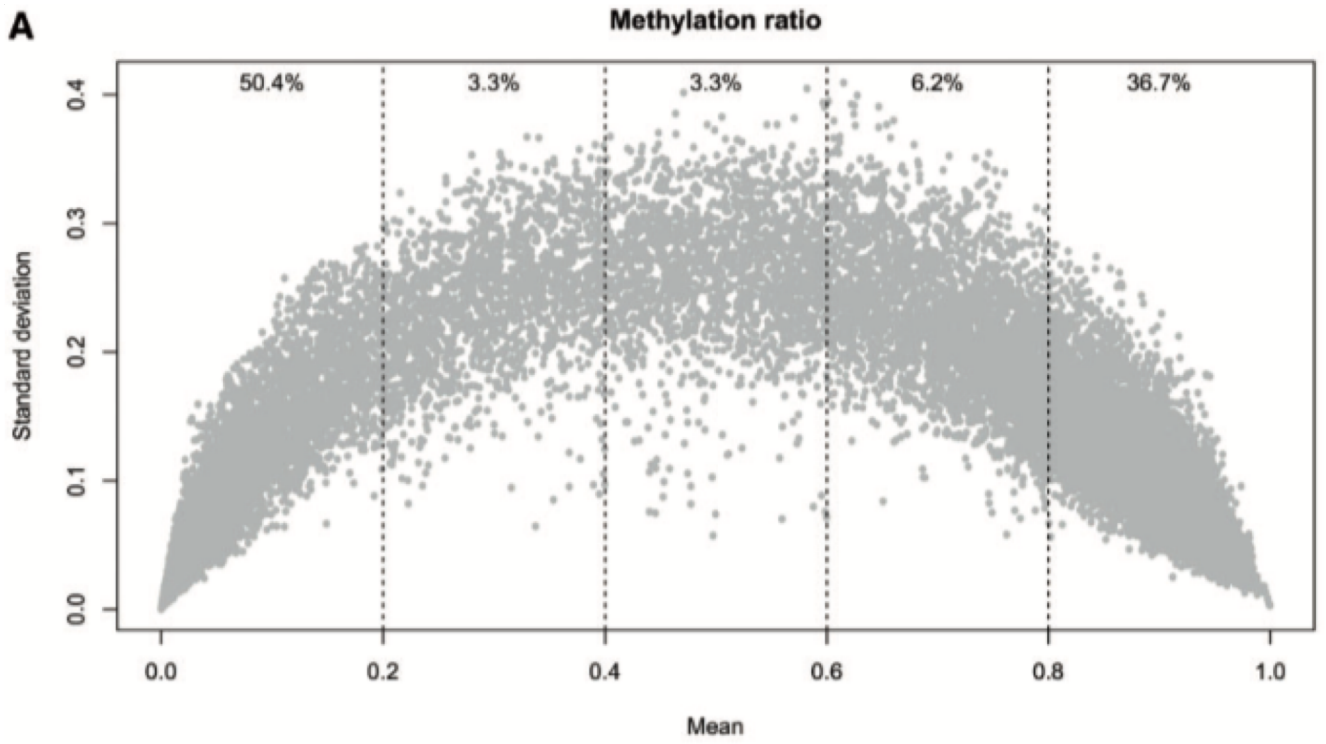
- Hierarchical model: Dirichlet-multinomial regression



La Rosa, Patricio S., et al. "Hypothesis testing and power calculations for taxonomic-based human microbiome data." *PloS one* 7.12 (2012): e52078.

Overdispersion – DNA methylation

- Hierarchical model: beta-binomial regression

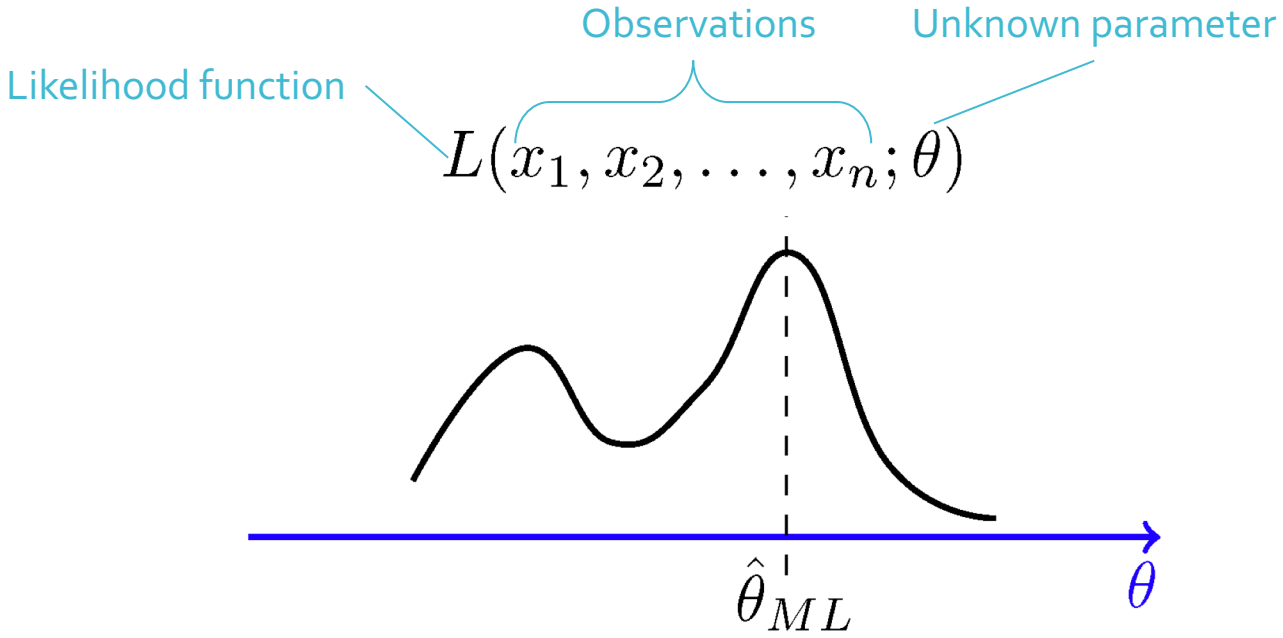


Zhang, Yun, Saurabh Baheti, and Zhifu Sun. "Statistical method evaluation for differentially methylated CpGs in base resolution next-generation DNA sequencing data." *Briefings in bioinformatics* 19.3 (2016): 374-386.

Model Fitting and Inference for GLM

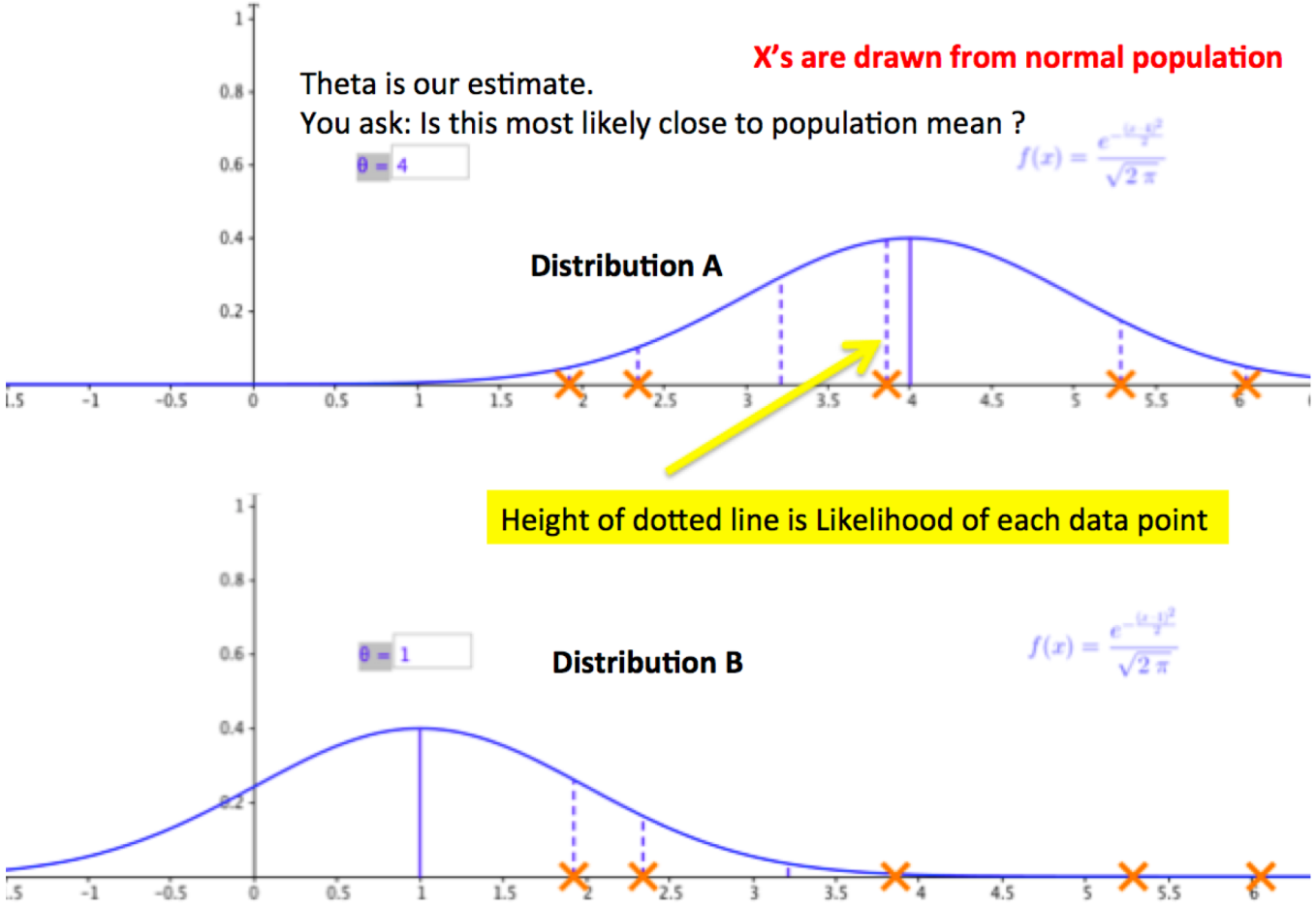
Fitting GLM

- GLM can be solved by maximum likelihood (ML) method
- Likelihood is a function of observations and unknown parameter
- Find $\hat{\theta}_{ML}$ that maximizes the likelihood function



- Using canonical link, $\theta = \boldsymbol{x}^T\boldsymbol{\beta} \Rightarrow \hat{\theta} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ ("easy" computation)

Maximum Likelihood Estimator (MLE)



- Maximum likelihood estimator picks the distribution that has the highest sum of dotted lines

Inference

- Analogies between linear regression and GLM

	Linear regression	GLM
Variance partitioning	Residual "sum of squares"	Deviance
Significance of coefficient	t-test	Wald test
Significance of model	F-test	Likelihood ratio test

- Deviance

$$D = -2 \ln \frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}}$$

- Wald test

$$W^2 = \left(\frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \right)^2 \sim \chi_1^2 \Rightarrow W = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \sim N(0,1)$$

It looks like a t-test; and W is sometime reported as t-statistic

- Likelihood ratio test

$$D_{\text{null}} - D_{\text{fitted}} = -2 \ln \frac{\text{likelihood of the null model}}{\text{likelihood of the fitted model}} \sim \chi^2_{\text{df}}$$

Derivation based on large sample theory

Final Comments

- Standard linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon, \quad \text{where } \epsilon \sim N(0, \sigma^2) \text{ i.i.d.}$$

- Assumptions of linear regression: LINE
  - The response Y and covariate x have Linear relationship
  - The errors are Independent
  - The errors are Normally distributed
  - The errors have Equal variances ("homoscedasticity")

- Generalized linear model (GLM)
  - If errors are *not* normally distributed
- Linear mixed-effect regression model (LMER) ← Next topic
  - If errors are *not* independent (i.e. correlated)