High-Throughput Sequencing Course
DESeq Model for RNA-Seq

Biostatistics and Bioinformatics

Summer 2019
Outline

- Review: Standard linear regression model (e.g., to model gene expression as function of an experimental condition or continuous covariate)
- Review: Logistic model: To model probability of a binary event as a function of a covariate
- Parameter interpretation: Linear and logistic regression
- Introduction: Negative binomial regression model for RNA-Seq
- Overview: Maximum likelihood estimation
Linear Regression Example: Gene Expression

- Consider the simple linear regression model

\[ Y = \beta_0 + \beta_1 x + \epsilon, \]

where

- \( x = 0 \) (untreated)
- or \( x = 1 \) (treated)

- \( Y \) is the observed "expression" of the gene
- \( \epsilon \) is the measurement noise term
- We assume that it follows a normal distribution with mean 0 and variance \( \sigma^2 \)
**Reminder: Important Fact about Normal Distribution**

- Consider a normal distribution with mean 0 and standard deviation $\sigma$.
- If the data are shifted by a constant $\mu$, then
  1. resulting distribution remains normal
  2. The mean of the new distribution is $\mu + 0 = \mu$
  3. Its standard deviation remains unchanged
- The last two (but not first) property are true for any distribution.
- Recall $Y = \beta_0 + \beta_1 x + \epsilon$
- $Y$ follows a normal distribution with mean $\mu = \beta_0 + \beta_1 x$
  and variance $\sigma^2$
- **IMPORTANT:** $\mu$ depends on $x$ (unless of course $\beta_1 = 0$)
Linear Regression Example: Interpretation

▶ Model

\[ Y = \beta_0 + \beta_1 x + \epsilon, \]

▶ The goal of (mean) regression is to estimate the expected value of \( Y \) given treatment status

▶ Conditional on \( x = 0 \) (i.e., not receiving treatment), the expected value of \( Y \) is

\[ \beta_0 + \beta_1 \times 0 = \beta_0 \]

▶ Conditional on \( x = 1 \) (i.e., receiving treatment), the expected value of \( Y \) is

\[ \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1 \]
General Conditional Expectation

- Expectation is another word for average
- We can write the conditional expectation of $Y$ given that $X = x$ as $E[Y|X = x]$
- English: This is the average value of the outcome $Y$ if the value of $X$ is equal to $x$
- The unconditional expectation of $Y$ is denoted by $E[Y]$
- If $Y$ does not depend on $X$, then $E[Y|X = x] = E[Y]$ for every $x$
- The goal of linear regression is to model $E[Y|X = x]$ as ”Linear” function
- Our Example: $E[Y|X = x] = \beta_0 + \beta_1 x$
**Linear Regression Example: Interpretation**

- **Model**
  
  \[ Y = \beta_0 + \beta_1 x + \epsilon, \]

- \( \beta_0 \) (the intercept) is the expected value of \( Y \) if no treatment is administered (average baseline value)

- \( \beta_1 \) is the treatment effect

- If treatment is administered, the expected value of expression is
  - increased by \( \beta_1 \) units if \( \beta_1 > 0 \)
  - decreased by \( \beta_1 \) units if \( \beta_1 < 0 \)
  - unchanged if \( \beta_1 = 0 \)
**Linear Regression Example: Continuous Covariate**

- **Model**

  \[ Y = \beta_0 + \beta_1 x + \epsilon, \]

  where \( x \) is continuous (quantitative)

- **If** \( \beta_1 > 0 \), then increasing \( x \) by one unit, increases \( Y \) on average by \( \beta_1 \) units

- **If** \( \beta_1 < 0 \), then increasing \( x \) by one unit, decreases \( Y \) on average by \( \beta_1 \) units

- **If** \( \beta_1 = 0 \), then changes in \( x \) do not affect the expected value of \( Y \)
Regression for Binary Outcomes

▶ Suppose that $Y$ is a binary outcome
▶ It assumes values 0 or 1
▶ This is a count outcome
▶ Consider the previous model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

▶ Is it appropriate? Why or why not?
Logistic Regression

- Relate the probability of the outcome of the event $Y = 1$ to treatment
- More specifically, relate the log-odds to the treatment
- The log-odds will be modeled as a linear function of $x$
  \[ \beta_0 + \beta_1 x + \epsilon \]

- This is an example of a generalized linear model (GLM)
- Note: The model used by DESeq is a GLM on the basis of the NB (instead of binomial distribution)
- The expected outcome of $Y$ is not modeled directly as a linear function
- A transformation of the expected outcome of $Y$ is modeled as a linear function
Expected value of a binary event

- Suppose that $Y$ assumes 1 with probability $\pi$ or 0 with probability $1 - \pi$
- $P(Y = 1) = \pi$ and $P(Y = 0) = 1 - \pi$
- **IMPORTANT:** $P(Y = 1) = E(Y)$
- The expected value of $Y$ is the probability that it assumes the value 1
- Why?
Relationship between $x$ and $\frac{\exp(x)}{1+\exp(x)}$
Odds vs Probability

- Suppose that $\pi = P(Y = 1)$
- The odds of the event $Y = 1$ (to occur) is defined as

$$\text{Odds}[Y = 1] = \frac{\text{Probability that } Y = 1 \text{ occurs}}{\text{Probability that } Y = 1 \text{ does not occur}} = \frac{\pi}{1 - \pi}$$
Odds Ratio Versus Relative Risk

▶ $\pi_0 = P[Y = 1|X = 0]$: Probability that the event occurs if sample is not treated

▶ $\pi_1 = P[Y = 1|X = 1]$: Probability that the event occurs if $X = 1$ sample is treated

▶ The odds-ratio is

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}$$

▶ The relative risk is

$$RR = \frac{\pi_1}{\pi_0}$$
The Logistic Model

- The log-odds of the event $Y = 1$

$$\log \frac{P(Y = 1|X = x)}{1 - P(Y = 1|X = x)} = \beta_0 + \beta_1 x$$

- or equivalently

$$\log \frac{E(Y|X = x)}{1 - E(Y|X = x)} = \beta_0 + \beta_1 x$$

- or equivalently

$$P(Y = 1|X = x) = E(Y|X = x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$
Parameter Interpretation

- If $\beta_1 > 0$, a unit increase in $x$, results in an expected increase of $\exp(\beta_1)$ in the odds of the event
- If $\beta_1 < 0$, a unit increase in $x$, results in an expected decrease of $\exp(\beta_1)$ in the odds of the event
- If $\beta_1 = 0$, then changes in $x$ do not affect the odds of realization of the event
For a probability $\pi$, define the "logit" transformation as

$$\log \frac{\pi}{1 - \pi}$$

This is the log-odds of an event with probability $\pi$

Note that in the logistic model, the probability of the event is linear in the parameter through this logit transformation

$$\log \frac{E(Y|X = x)}{1 - E(Y|X = x)} = \beta_0 + \beta_1 x$$

In the GLM literature, this is called the link function
**OVERDISPERSION**

- Recall that if $K$ follows a binomial distribution with parameters $n$ and $\pi$, then
  - mean $\mu = n\pi$
  - variance $\sigma^2 = n\pi(1 - \pi)$

- Clustering in the data results in the actual variance to be different than the nominal variance ($n\pi(1 - \pi)$)
  - Overdispersion: Actual variance is larger than nominal variance
  - Underdispersion: Actual variance is smaller than nominal variance

- The choice of a GLM and evaluation of its performance should start and end with considering/addressing the overdispersion issue

- The use of Poisson (actually a variation thereof) and Negative Binomial models are two common choices for GLM for overdispersed data
**Generalized Linear Models (GLM)**

Define \( \mu_x = E(Y|X = x) \) as the expected value of the outcome given treatment status \( (x = 0 \text{ or } x = 1) \)

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Link</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial</td>
<td>( 0, 1, \ldots, n )</td>
<td>( \beta_0 + \beta_1 x = \log \frac{\mu_x}{1-\mu_x} ) ( \mu_x = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} )</td>
</tr>
<tr>
<td>Poisson</td>
<td>( 0, 1, 2, \ldots )</td>
<td>( \beta_0 + \beta_1 x = \log(\mu_x) ) ( \mu_x = \exp(\beta_0 + \beta_1 x) )</td>
</tr>
<tr>
<td>Negative Binomial</td>
<td>( 0, 1, 2, \ldots )</td>
<td>( \beta_0 + \beta_1 x = \log(\mu_x) ) ( \mu_x = \exp(\beta_0 + \beta_1 x) )</td>
</tr>
</tbody>
</table>
GENERAL NOTE

- Recall the simple linear regression model for expression

\[ Y = \beta_0 + \beta_1 x + \epsilon, \]

where
- \( x = 0 \) (untreated)
- or \( x = 1 \) (treated)

- \( Y \) is the observed ”expression” of the gene
- \( \epsilon \) is the measurement noise term
- The parameter of interest is \( \beta_1 \) (the treatment effect)
- There are two other unknown parameters, \( \beta_0 \) and \( \sigma^2 \) the estimation procedure has to deal with in a principled manner
- \( \beta_0 \) and \( \sigma^2 \) are nuisance parameters
- They are not of primary (or any) interest. But you have to deal with them!
General Hypothesis

- Is the RNA abundance level for any of the \(m\) genes affected by treatment
- Let \(H_j\) denote the null hypothesis for gene \(j\)
- \(H_j\): The RNA abundance level for gene \(j\) is not affected by treatment
- \(\bar{H}_j\): The RNA abundance level for gene \(j\) is affected by treatment
- The global null hypothesis: \(H_1\) and \(H_2\) and .... and \(H_m\) are all true
- The global alternative: \(\bar{H}_1\) or \(\bar{H}_2\) or .... or \(\bar{H}_m\) is true
- In other words, under the alternative at least one of the marginal null hypotheses is false
**Observed Data**

- Some notation
  - $n$ denotes the number of samples
  - $m$ denotes the number of genes
  - $K_{ij}$ denotes the observed number of reads mapped to gene $i$ for sample $j$
  - $x_j = 0$ or $1$ denotes the treatment status for sample $j$

- What is observed for sample $j$ is the vector

$$K_{1j}, \ldots, K_{mj}, x_j$$

- In other words $m$ counts (one per gene) and the experimental factor

- Note that the $K_{ij}$ form a table of counts of dimension $n \times m$ ($n$ samples and $m$ genes)
**DESeq: Notation for Negative Binomial Distribution**

- The count $K$ is assumed to follow a negative binomial distribution with parameters $p \in (0, 1)$ and $r > 1$
- The distribution is PMF is

$$P(K = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k,$$

for $k = r, r + 1, \ldots$

- Rather than considering the model as NB$[p, r]$ we will consider it as NB$[\mu, \alpha]$, where

$$P[K = k] = \frac{\Gamma[k + \alpha^{-1}]}{\Gamma[\alpha^{-1}]} \left( \frac{1}{1 + \mu \alpha} \right)^{\alpha^{-1}} \left( \frac{\mu}{\alpha^{-1} + \mu} \right)^k,$$

where $k = 0, 1, \ldots$
DESeq: Notation

- $K_{ij}$ denotes the observed number of reads mapped to gene $i$ for sample $j$
- $K_{ij}$ follows a negative binomial distribution with
  - Mean $\mu_{ij}$ (indexed by gene $i$ and sample $j$)
  - Dispersion parameter $\alpha_i$ (indexed by the gene $i$)
- The mean is assumed to be $\mu_{ij} = s_j q_{ij}$ where
  - $\log q_{ij} = \beta_{i0} + \beta_{i1} x_j$
  - $s_j$ is a gene $j$ specific normalization constant
**DESeq: Reformulate Hypotheses**

- **Hypotheses of interest**
  - The global null hypothesis: \( H_1 \) and \( H_2 \) and .... and \( H_m \) are all true
  - The global alternative: \( \bar{H}_1 \) or \( \bar{H}_2 \) or .... or \( \bar{H}_m \) is true

- **Reformulation**
  - The global null hypothesis: \( \beta_{11} = 0 \) and \( \beta_{21} = 0 \) and .... and \( \beta_{m1} = 0 \)
  - In other words, all of the \( \beta_{j1} \) are equal to zero
  - The global alternative: \( \beta_{11} \neq 0 \) or \( \beta_{21} = 0 \) or .... or \( \beta_{m1} = 0 \)
  - In other words, at least one of the \( \beta_{j1} \) is not equal to zero
$K_{ij}$ follows a negative binomial distribution with mean $\mu$ and dispersion parameter $\alpha$. 
**DESeq: Assumption on Mean of Distribution**

- Conditional on the treatment status of sample $j$ ($x_j = 0$ or 1), the expected value of $K_{ij}$ is

$$\mu_{ij} = s_j \times q_{ij}$$

where

$$\log q_{ij} = \beta_{i0} + \beta_{i1} x_j$$

- Note that two regression parameters are indexed by $i$
- Why? Because these are gene $i$ specific parameters
- Why is $x_j$ not indexed by $i$?
- Final Assumption: $s_{ij} = s_j$
- In other words: Within sample $j$, the normalization parameter is constant across the genes
- How many assumptions so far?
DESeq: Main Parameters and Nuisance Parameters

- The $m$ main parameters of interest
  
  \[ \beta_{11}, \ldots, \beta_{m1} \]

- The unknown nuisance parameters are
  - The $m$ gene specific intercepts
    
    \[ \beta_{10}, \ldots, \beta_{m0} \]
  
  - the $n$ sample specific normalization constants
    
    \[ s_1, \ldots, s_n \]
  
  - The $m$ gene specific nuisance parameters
    
    \[ \alpha_1, \ldots, \alpha_m \]
**DESeq: Main parameters and Nuisance Parameters**

- Assuming the model assumptions are correct, the estimation of the regression parameters $\beta_{i0}, \beta_{i1}$ is fairly straightforward
- The DESeq authors propose to estimate the normalization constant for sample $j$ as

$$s_j = \text{median} \frac{K_{ij}}{K_i^R},$$

where

$$K_i^R = \left( \prod_{j=1}^{m} K_{ij} \right)^{\frac{1}{m}}$$

- Here $K_i^R$ is the geometric mean of $K_{i1}, \ldots, K_{in}$ (the $n$ counts for gene $i$)
- The median is taken over all $m$ genes for which $K_i^R$ is positive
**DESeq: Dispersion Parameter**

- A key issue in using the NB model is proper handling of the gene specific dispersion parameters \( \alpha_1, \ldots, \alpha_m \)

- The estimation of the dispersion parameter is a challenging task

- DESeq2 assumes that \( \alpha_i \) is random following a normal distribution

- The results are sensitive to the estimates

- One of the key differences between DESeq2 and DESeq is the approach taken to estimate these nuisance parameters
DESeq Software Overview

- The analysis of RNA-Seq data using the DESeq2 package will be reviewed in detail in the upcoming weeks.
- The estimation and inference for the model is done through the DESeq function.
- It performs the following steps in the order given:
  1. estimation of size factors $s_1, \ldots, s_n$
  2. estimation of dispersion parameters $\alpha_1, \ldots, \alpha_m$
  3. Fit NB GLM model.
DESeq: Model Exercise

- $K_{ij}$ denotes the observed number of reads mapped to gene $i$ for sample $j$
- $x_j = 0$ or $1$ denotes the treatment status for sample $j$
- Say we want to account for another covariate $z_j$ (e.g., temperature)
- What is observed for sample $j$ is the vector $K_{1j}, \ldots, K_{mj}, x_j, z_j$

Questions
- State the hypotheses
- Propose a model (that incorporates the additional covariate)
- List any assumptions that you have made
**DESeq: Model Exercise**

- The null hypothesis
  
  \[ H_0 : \beta_{11} = 0 \text{ and } \beta_{21} = 0 \text{ and } \ldots \beta_{m1} = 0 \]

- Conditional on \( x_j \) and \( z_j \), the observed number of reads mapped to gene \( i \) for sample \( j \), \( K_{ij} \), follows a negative binomial distribution with
  
  - Mean \( \mu_{ij} \)
  
  - Dispersion parameter \( \alpha_i \) (gene specific)

- Conditional on the treatment status of sample \( j \) (\( x_j = 0 \) or 1) and the temperature \( z_j \), the expected value of \( K_{ij} \) is

  \[ \mu_{ij} = s_j \times q_{ij} \]

  where

  \[ \log q_{ij} = \beta_{i0} + \beta_{i1} x_j + \beta_{i2} z_j \]

- The normalization parameters are assumed to be sample (not gene) specific (\( s_{ij} = s_j \))
DESeq: Model Nuisance Parameter

- The $m$ main parameters of interest

$$\beta_{11}, \ldots, \beta_{m1}$$

- The unknown nuisance parameters are
  - The $m$ gene specific intercepts

$$\beta_{10}, \ldots, \beta_{m0}$$

  - The $m$ gene specific coefficients for the new covariate

$$\beta_{12}, \ldots, \beta_{m2}$$

- The $n$ sample specific normalization constants

$$s_1, \ldots, s_n$$

- The $m$ gene specific nuisance parameters

$$\alpha_1, \ldots, \alpha_m$$
Edger: Another NB Model for RNA-Seq Counts

- Assume that the $K_{ij}$ follows a NB distribution with mean $\mu_{ij}$ and dispersion parameter $\alpha_i$
- The mean (conditional on treatment status $x$) is

$$\mu_{ij} = M_j p_{xi}$$

where

- $M_j$ is the library size (total number of reads for sample $j$
- $p_{xi}$ is the relative abundance of the gene $i$ given treatment status $x$
  - $p_{0i}$ is the relative abundance of the gene $i$ given no treatment
  - $p_{1i}$ is the relative abundance of the gene $i$ given treatment

- Treatment changes the abundance of RNA in gene $i$ if $p_{0i} \neq p_{1i}$
- This is same distributional assumption as in DESeq
MLE ILLUSTRATION

- In a GLM, the parameters $\beta_{i0}$ and $\beta_{i1}$ are estimated using the method of Maximum likelihood (MLE)
- We illustrate the method using this coin tossing example:
- We toss a coin once and record the number of heads
- Suppose that you conduct two independent replicates of this experiment
- $K_1$ the number of events (among $n = 1$ trial) in experiment 1
- $K_2$ the number of events (among $n = 1$ trial) in experiment 2
- The PMF of $K_1$ is
  \[ P(K_1 = k) = \pi^k (1 - \pi)^{1-k} \]
- The PMF of $K_2$ is
  \[ P(K_2 = k) = \pi^k (1 - \pi)^{1-k} \]
- Here $k = 0$ or 1
**Joint Distribution**

- \( P(K_1 = k_1) \) denotes the probability of the event that \( K_1 = k_1 \)
- \( P(K_2 = k_2) \) denotes the probability of the event that \( K_2 = k_2 \)
- These are called marginal probabilities
- What is \( P(K_1 = k_1, K_2 = k_2) \)?
- This is probability of the event that \( K_1 = k_1 \) and \( K_2 = k_2 \)
- If you assume that these are independent tosses then
- \( P(K_1 = k_1, K_2 = k_2) = P(K_1 = k_1) \times P(K_2 = k_2) \)
- In other words, the probability of the joint event is equal to the probability of the marginal events.
Likelihood

- Suppose that the realized value of $K_1$ is $k_1$
- Unlike $K_1$, $k_1$ is a fixed non-random number
- The likelihood of $\pi$ given the observed data $k_1, k_2$ is
  \[
  L(\pi) = \pi^{k_1} (1 - \pi)^{1-k_1} \pi^{k_2} (1 - \pi)^{1-k_2}
  \]
- Note that this is the joint probability of the events evaluated at the realized values
Repeat the experiment $B$ times

The joint PMF is

$$P(K_1 = k_1, \ldots, K_B = k_B) = \pi^{k_1}(1-\pi)^{1-k_1} \times \ldots \times \pi^{k_B}(1-\pi)^{1-k_B}$$

Note that the implicit assumption is that the experiments are mutually independent

Under this assumption, the joint PMF is the product of the marginal PMFs

Plugging in the observed counts into the joint PMF yields the likelihood function
**Binomial Example: Observed data**

- Observed data $x_1 = 1, x_1 = 0, x_3 = 0, x_4 = 0$ and $x_5 = 1$
- What is the likelihood?

```r
set.seed(2131)
x = rbinom(5, 1, 0.5)
x
```
```r
## [1] 1 0 0 0 1
```
Binomial Example: Likelihood

- Observed data $x_1 = 1, x_1 = 0, x_3 = 0, x_4 = 0$ and $x_5 = 1$
- The likelihood

$$L[\pi] = \pi^{x_1}(1 - \pi)^{x_1} \times \pi^{x_2}(1 - \pi)^{x_2} \times \pi^{x_3}(1 - \pi)^{x_3} \times \pi^{x_4}(1 - \pi)^{x_4} \times \pi^{x_5}(1 - \pi)^{x_5} \times$$

$$= \pi^1(1 - \pi)^{1-1} \times \pi^0(1 - \pi)^{1-0} \times \pi^0(1 - \pi)^{1-0} \times \pi^0(1 - \pi)^{1-0} \times$$

$$= \pi^2(1 - \pi)^3$$

- Given the observed data find the value of $\pi$ that maximizes this probability
Binomial Example: Maximum Likelihood

The maximum value of the function $L[\pi] = \pi^2 (1 - \pi)^3$ occurs at $\pi = 0.4$. 

![Graph showing the maximum likelihood function $L(\pi)$ for the binomial distribution with $\pi = 0.4$.](graph.png)
Maximum Likelihood Calculation for NB

- For gene $i$, let $k_{i1}, \ldots, k_{in}$ the $n$ observed counts
- For patient $j$ plug the observed count $k_{ij}$ into the PMF of the NB distribution $f[k_{ij}; \mu_{ij}; \alpha_i]$
- Write the likelihood function as a product of these $n$ terms

$$L = \prod_{j=1}^{n} f[k_{ij}; \mu_{ij}; \alpha_i] = f[k_{ij}; \beta_{0i}, \beta_{1i}, s_j, \alpha_i]$$

- The function depends on $\beta_{0i}, \beta_{1i}, s_j$ and $\alpha_i$
- One approach: Come up with some estimates of $s_j$ and $\alpha_i$ and plug them into the likelihood
- Pretend that these are the true values
- Now the likelihood is only a function of $\beta_{0i}$ and $\beta_{1i}$