Lecture 7: Statistics For RNA-Seq

BIOINF3005/7160: Transcriptomics Applications

Dr Stephen Pederson

Bioinformatics Hub,
The University of Adelaide

May 4th, 2020

Recap of Continuous Data

Continuous Variables

- Continuous variables can take any value on the number line
  - i.e $-\infty < x < \infty$
- Are unbound at either limit
- For analysis, being continuous within the complete range of possible values is enough
- Microarray fluouresence intensities are bound at both extrema:
  - $PM \geq 0$ and $PM \leq 2^{16}$

Normally Distributed Data

- Normally distributed data **must** be continuous
  - If not, the bell-curve becomes discrete
  - Boundary points can also be problematic, but Truncated-Normal distributions exist.
- T-tests rely on the assumption of Normality
- Linear Regression also relies on the assumption of Normality

Linear Regression

- Our recent pracs have all been fitting linear regression models
- We attempt to fit a line through our response (y) and predictor (x) variables
- The interpretation is always:

For a 1 unit increase in predictor x, we expect to see '...' change in our response variable y

Linear Regression

- Our model coefficients provide this estimate of change
- Represent the slope of the line
- *Predictor variables* can be discrete but *response variables* must be continuous
    - e.g. sample groups are a common discrete predictor

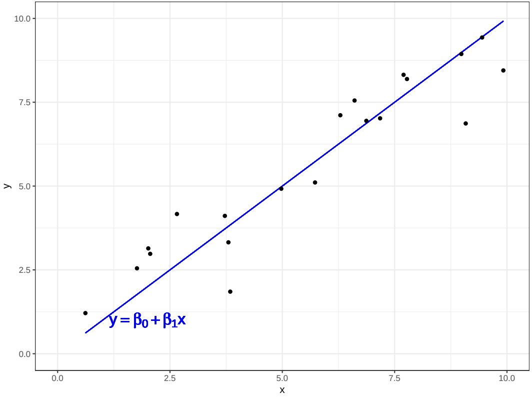
Linear Regression

For observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ and predictors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$

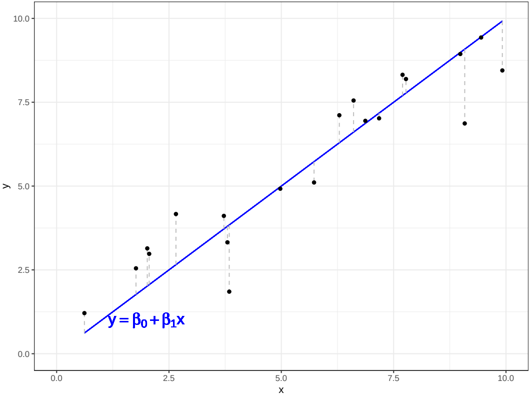
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- We are trying to fit a straight line with intercept (β_0) and slope (β_1)
- Points never line up exactly on the line, so we need an error term ε
- The error term (also called *residuals*) is $\varepsilon_i \sim \mathcal{N}(0, \sigma)$

Linear Regression



Linear Regression



Linear Regression

The four assumptions of linear regression

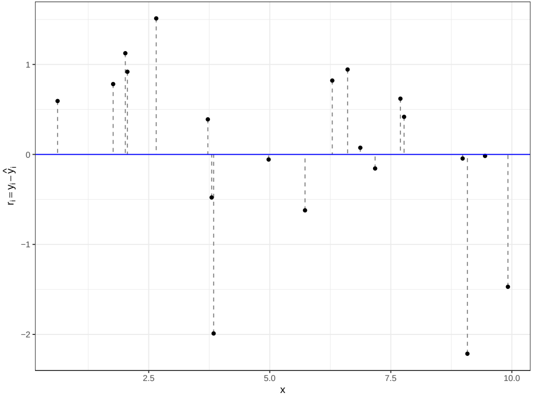
1. Normality: Residuals are normally distributed
2. Homoscedasticity: Variance is constant across the range of the data
3. Linearity: Data is linear
4. Observations are independent

Linear Regression

All of these relate to the residuals: $\varepsilon_i \sim \mathcal{N}(0, \sigma)$

1. Normality: $\mathcal{N}$
2. Homoscedasticity: σ
3. Linearity: $\mu = 0$
4. Observations are independent: (We just assume, unless we know better)

Linear Regression



Linear Regression

- To fit a linear regression model we use least squares

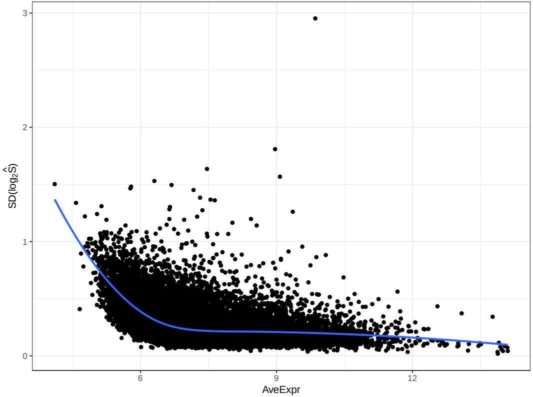
$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- Then we check residuals for all assumptions (linearity, normality, constant variance)

Microarrays

- All of the above works well for microarrays
- Boundary points of 0 & 2^{16} are ignored
- We know variance is connected to the fluorescence intensity
- Using $y = \log_2 \hat{S}$ gives almost constant variance
  - Become important when genes are DE
- We can't check assumptions for every gene, but generally they hold

Microarrays



Discrete Distributions

Discrete Data

- Discrete data involves different types of measurements than continuous
- Common types of counts:
  1. Number of successes in a binary test e.g. number of 6's rolled
  2. Number of events in a fixed unit of measurement, e.g. cars passing per minute

Discrete Data

- Discrete data involves different types of measurements than continuous
- Common types of counts:
    1. Number of successes in a binary test e.g. number of 6's rolled
    2. Number of events in a fixed unit of measurement, e.g. cars passing per minute
- The number of successes (1) can be modelled using the *Binomial* and *Hypergeometric* distributions
- The number of events (2) can be modelled using the *Poisson* and *Negative-Binomial* distributions

Binomial Data

- If we have a bag of 100 balls: 20 red and 80 blue
- The probability of success (i.e grabbing a red ball) is $\pi = 0.2$
- This is the classic binomial scenario
- If we return the ball we've just taken $\implies \pi = 0.2$.
- What if we keep the ball and don't replace it?

Hypergeometric Data

- The Hypergeometric Distribution is what happens when we sample **without replacement**
- The most common representation of this is a 2 × 2 table
- The most appropriate test is Fisher's Exact Test

Fisher's Exact Test

- Developed by RA Fisher (prior to arrival at Adelaide)
- H_0: *No association between variables* Vs H_A: *Some association between variables*

Fisher's Exact Test

- Developed by RA Fisher (prior to arrival at Adelaide)
- H_0: *No association between variables* Vs H_A: *Some association between variables*

	DE	**not DE**	n
On Chr1	100	900	1000
Not on Chr1	1000	19000	20000

Here 1 in 11 DE genes is on Chr1, whilst 1 in $\sim$22 *not DE* genes are on Chr1

Fisher's Exact Test

- Was there an association? ($p = 3.74 \times 10^{-10}$)
- The test is two-sided: i.e. association can be in either direction
- Note that once we've sampled a gene, it can't be replaced $\implies$ *hypergeometric*

Fisher's Exact Test

- Was there an association? ($p = 3.74 \times 10^{-10}$)
- The test is two-sided: i.e. association can be in either direction
- Note that once we've sampled a gene, it can't be replaced $\Longrightarrow$ *hypergeometric*
- We use this for enrichment testing (next week) and a variation is used for Differential Expression in RNA Seq
- Notice there is no provision for replicates within groups under this layout

Poisson Distributed Data

- *Poisson* distributed data is based on a rate of occurrence, e.g. mobile phone networks
- We have a count *per fixed unit* $\implies$ a rate is involved
- The rate parameter (λ) is the average number of number of events / unit
- **The standard deviation is the same as the rate**
- This is fundamentally different to the Normal Distribution $\implies$ μ and σ are independent

Poisson Distributed Data

```r
rpois(n = 1000, lambda = 1)
#   0   1   2   3   4   5
# 390 328 202  57  20   3

```

Poisson Regression

- To fit Poisson-distributed data, we use *Generalised Linear Models* (GLM)
- Poisson GLMs are sometimes known as *log-linear* models
- The formula looks the same in R, but is fitted using **Maximum Likelihood** not Least Squares
- The response value is fitted on the (natural) log-scale
- Errors are no longer normally distributed (should be Poisson)
- `glm(y ~ predictors, family = "poisson")`

Poisson Regression

- The default formula relies in the fixed unit being identical
- What if we're counting trees for multiple species across multiple forests
- The forest size always changes & the species distribution changes within each forest
- We can supply an offset term to the model which accounts for this
    - Effectively standardises the unit of measurement

Negative Binomial Data

- What happens when our ‘real-world’ data is more variable?
  - When variance is clearly $> \lambda$
- This is known as over-dispersed data
- Best fit using a Negative Binomial model as a GLM
- Essentially the same as a Poisson but with more wiggle room

Applications to RNA Seq

RNA Seq Libraries

- After aligning to a reference and counting reads for each gene $\Longrightarrow$ gene-level counts
- We face many familiar issues:
  - How do we normalise the data?
  - How do we test for Differential Expression?
- We generally refer to each set of counts as a ‘library’
- Library Sizes are a big issue in RNA Seq

Normalisation

- How do we adjust for library size differences
- Some libraries amplify well/poorly?
- How does this affect library composition?
- Do some highly expressed genes ‘dominate’ a library?
- Was there a different response to GC content across individual libraries?
- Longer genes will also receive more counts

Unlike microarrays, *we don’t adjust our counts directly*!

Normalisation

- Essentially we're fitting the rate of observing counts in each library
- This is impacted by total library size
- RNA-Seq data tends to be analysed using Negative Binomial Models
    - Data is over-dispersed (i.e. more variable) than a Poisson model
- We can use the *offset* trick we introduced earlier

Normalisation

- A naive approach would be to adjust for library sizes
- The most common strategy is Trimmed Mean of M-values (TMM)
  - Effectively adjusts for distortions in library composition and total size
- Can also use Conditional-Quantile Normlisation (CQN)
  - This adjusts for sample-specific GC effects and/or sample-specific length effects

TMM Normalisation

- Here we use our M and A values again
- Assumes that most genes are not differentially expressed
- For a pair of samples $(k = 1, 2)$ and a given gene g with counts y_{gk}

$$M_g = \log_2 \frac{y_{g1}/N_1}{y_{g2}/N_2}$$

$$A_g = \frac{1}{2}\left(\frac{y_{g1}}{N_1} + \frac{y_{g2}}{N_2}\right)$$

- In all of the above N_k represents the total library size

TMM Normalisation

- Data across all genes is trimmed 30% (M-values) and 5% (A-values)
- The sum of the weighted trimmed M-values is then calculated
- A *sample-level* normalisation factor is calculated by comparing to a reference sample
- This value is provided to the model as an offset

CQ Normalisation

- GC content and gene length doesn’t change across samples
- Sometimes libraries respond differently during library preparation
- May be PCR-related or fragmentation-related
- CQN provides an *gene-level* and *sample-level* offset

Dispersions

- Under the NB model there is Poisson variability (μ) + overdispersion (φ)
- $E(Y) = \mu$ and $\mathrm{Var}(Y) = \mu + \varphi\mu^2$, with $\varphi > 0$
- This is estimated as an overall value, and a gene-level value
- Specifically designed to handle unequal library sizes
  - $\hat{\varphi}$ is estimated using a quantile Conditional Maximum Likelihood model (qCML)[1]
  - The qCML procedure requires calculation of pseudo-counts, or pseudo-data

[1] Mark D. Robinson and Gordon K. Smyth. "Small-sample estimation of negative binomial dispersion, with applications to SAGE data". In: *Biostatistics* 9.2 (Aug. 2007), pp. 321–332. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxm030. eprint: https://academic.oup.com/biostatistics/article-pdf/9/2/321/17734659/kxm030.pdf. URL: https://doi.org/10.1093/biostatistics/kxm030.

The Exact Test

- For a simple 2 group comparison, an Exact Test can be used
- By using the qCML estimates, pseudo-data is identically distributed across samples
- Can be pooled within groups for Fisher's Exact Test[2]
- Provides a p-value for Differential Expression
  - $H_0 : \lambda_1 = \lambda_2$ with $H_A : \lambda_1 \neq \lambda_2$
  - logFC is effectively the $\Delta\lambda$ on the $\log_2$ scale

[2] Mark D. Robinson and Gordon K. Smyth. "Small-sample estimation of negative binomial dispersion, with applications to SAGE data". In: *Biostatistics* 9.2 (Aug. 2007), pp. 321–332. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxm030. eprint: https://academic.oup.com/biostatistics/article-pdf/9/2/321/17734659/kxm030.pdf. URL: https://doi.org/10.1093/biostatistics/kxm030.

GLM Approaches

- For more general approaches a GLM approach can be taken using a Negative Binomial as the underlying distribution
- Can fit simple 2-group comparisons or more complex designs
- Under these approaches, trended dispersions are used
- Analogous to moderated variances for the moderated T-test
  - A different Empirical Bayes model is used
  - Reduces false positives and false negatives simultaneously

GLM Approaches

- The effect of any predictor on the counts can be modelled
- The latest approach is to use the Quasi-Likelihood GLM fit (`glmQLFit()`)
- This has been shown to be the most reliable model
  - The original GLM models in `edgeR` don't strictly control the FDR[3]

[3] S. P. Lund et al. "Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates". In: *Stat Appl Genet Mol Biol* 11.5 (2012).

Differential Expression Testing

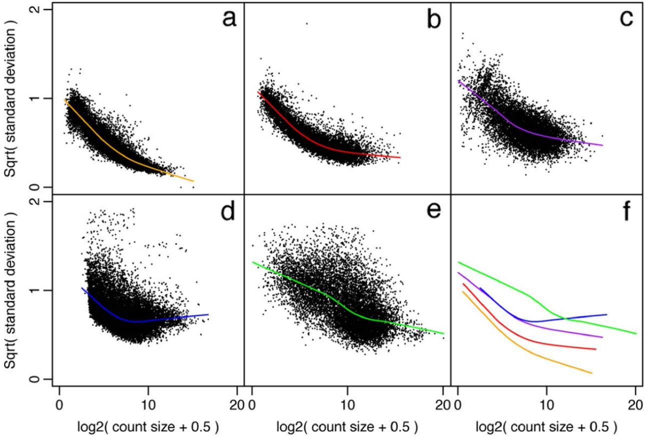
- For Microarray data, we perform a T-test on each model coefficient
- This is not possible under GLM approaches
- Instead we perform Likelihood Ratio Tests
  - Tests the 'Goodness of Fit' of two models
  - One with the model term, the other without
- For QL-GLM we can use a Quasi-Likelihood F-test
- Analogous to an ANOVA test

voom

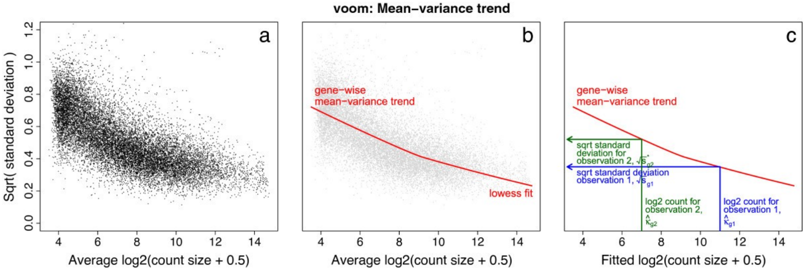
- An alternative to all of the above might be to transform the counts into continuous data
- How would we handle the mean-variance relationship?
- The voom approach is based on using Counts/Million, or logCPM[4]

[4] C. W. Law et al. "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts". In: *Genome Biol.* 15.2 (2014), R29.

voom


a
b
c
d
e
f
Sqrt(standard deviation)
Sqrt(standard deviation)
0
1
2
0
1
2
0
10
20
0
10
20
0
10
20
log2(count size + 0.5)
log2(count size + 0.5)
log2(count size + 0.5)

voom



- Predicted counts are obtained by fitting the CPM values
- Using the lowess curve based on counts, predicted standard deviations are obtained
- The inverse of predicted standard deviations (but squared) are the weights
- We can fit using limma and all assumptions of normality are back on the table

Why Normality?

- The suite of statistical tools available for Normal data is vary broad
- Weighted regression, Mixed-effects/Nested Models, T-tests
- Voom brings this back into play

A Final Word

- A common measure for gene expression is Counts / Million (CPM) or logCPM
- Very intuitive measure and useful for visualisation
- Only voom uses them for fitting data, by managing the mean-variance relationship
- Nearly all other models use the raw counts for fitting
  - Early approaches used RPKM and FPKM $\implies$ now discredited
  - Another common value is Transcripts Per Kilobase Million (TPM) $\implies$ incorporates gene length
- CPM is good across samples; TPM is good within samples