# Linear models for gene expression data analysis

Mikhail Dozmorov

2021-03-15

# General framework for differential expression

- Linear models
- Model the expression of each gene as a linear function of explanatory variables (Groups, Treatments, Combinations of groups and treatments, Etc. . . )

$$y = X\beta + \epsilon$$

- $y$ - vector of observed data
- $X$ - design matrix
- $\beta$ - vector of parameters to estimate

# Example of a design matrix

Normal sample x 2

Cancer Sample x 2

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

$\beta_1$ = normal log-expression

$\beta_2$ = cancer – wt

$E[y_1]=E[y_2]=\beta_1$

$E[y_3]=E[y_4]=\beta_1+\beta_2$

# Example of a design matrix

More examples

- 6 samples
- 2 groups + drug treatment
- Group and treatment effect are additive

$$y = X\beta + \epsilon$$

| Group1 | Group 2- Group 1 | Drug dose |
|---|---|---|
| 1 | 0 | 0.25 |
| 1 | 0 | 1 |
| 1 | 0 | 4 |
| 1 | 1 | 0.25 |
| 1 | 1 | 1 |
| 1 | 1 | 4 |

3 coefficients to estimate

# Linear model parameter estimation

Model is specified – how do we find the coefficients? $y = X\beta + \epsilon$

- Minimize squared error $\epsilon'\epsilon = (Y - X\beta)'(Y - X\beta)$
- Take derivative $\frac{d}{d\beta}((Y - X\beta)'(Y - X\beta)) = -2X'(Y - X\beta)$
- Set to 0, $-2X'(Y - X\beta) = 0$
- Solve $X'Y = (X'X)\beta$ $\beta = (X'X)^{-1}X'Y$

# Hypothesis testing

- Significance of coefficients is tested with a T-test

$\beta$ can be a vector. We can test the significance of any one coefficient $\beta_i$ via a T-test

$$t_{score} = \frac{\hat{\beta} - \beta_0}{SE_{\hat{\beta}}}$$

$$t_{score} = \frac{(\hat{\beta} - \beta_0)\sqrt{n-2}}{\sqrt{SSR / \sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$SSR = \sum_{i=1}^{n} \hat{\epsilon}^2$ - sum of squares of residuals, depends on the whole model

# Linear models and covariates

- Linear models are useful for including nuisance variables - technical factors
- Variables that have an effect on measurements but are not themselves of interest (e.g. sample storage time)
- Incorporating storage time gives smaller residuals and thus larger T-stats for the coefficient of interest

# Limma method

- Generalized the hierarchical model of Lonnstedt and Speed (2002) into a practical approach for general gene expression experiments.
- The model borrows information across genes to smooth out variances and uses posterior variances in a classical t-test setting.
- Completely data-dependent and uses empirical Bayes approach to estimate hyper parameters

# Limma method

Smyth et al. (2004) Statistical Applications in Genetics and Molecular Biology

- Uses a Bayesian hierarchical model in multiple regression setting.
- Borrows in formation from all genes to estimate gene specific variances.
- As a result, variance estimates will be "shrunk" toward the mean of all variances. So very small variance scenarios will be alleviated.
- Implemented in Bioconductor package "limma".

http://bioinf.wehi.edu.au/limma/

https://bioconductor.org/packages/release/bioc/html/limma.html

# Limma

- **design matrix**
  - defines which conditions arrays belong to
  - rows: arrays; columns: coefficients
- **contrast matrix**
  - specifies which comparisons you would like to make between the RNA samples
  - for very simple experiments, you may not need a contrast matrix
- Law, Charity W., Kathleen Zeglinski, Xueyi Dong, Monther Alhamdoosh, Gordon K. Smyth, and Matthew E. Ritchie. "A Guide to Creating Design Matrices for Gene Expression Experiments." F1000Research (December 10, 2020) - Design matrices for various experimental designs. Means model or mean-reference model.
- Soneson, C, F Marini, F Geier, MI Love, and MB Stadler. "ExploreModelMatrix: Interactive Exploration for Improved Understanding of Design Matrices and Linear Models in R" F1000Research, (June 4, 2020).

# Limma references


- Lönnstedt, Ingrid, and Terry Speed. "REPLICATED MICROARRAY DATA." Statistica Sinica 12, no. 1 (2002): 31–46. http://www.jstor.org/stable/24307034. - Empirical Bayes method for analyzing microarray replicates. Issues with simple approaches, proposed B statistics - a Bayes log posterior odds with two hyperparameters in the inverse gamma prior for the variances, and a hyperparameter in the normal prior of the nonzero means. Appendix - detailed definitions, derivations, and solutions.
- Smyth, Gordon K. "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." Statistical Applications in Genetics and Molecular Biology 3 (2004): Article3. doi:10.2202/1544-6115.1027. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.315.2066&rep=rep1&type=pdf - Linear models for differential analysis, moderated t-statistics via shrinkage of sample variance. Empirical estimation of Bayesian prior variance distribution and shrinkage hyperparameters.

# Extensions of Limma method

- Sartor, Maureen A., Craig R. Tomlinson, Scott C. Wesselkamper, Siva Sivaganesan, George D. Leikauf, and Mario Medvedovic. "Intensity-Based Hierarchical Bayes Method Improves Testing for Differentially Expressed Genes in Microarray Experiments." BMC Bioinformatics 7 (December 19, 2006): 538. https://doi.org/10.1186/1471-2105-7-538. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-538 - Intensity-Based Moderated T-statistic (IBMT). Empirical Bayes approach allowing for the relationship between variance and gene signal intensity (estimated with loess). Brief description of previous methods (Smyth, Cyber-T). Details of Smyth hierarchical model and moderated t-statistics, estimation of hyperparameters with implementation of variance-signal. Software at http://eh3.uc.edu/ibmt/.
- Lianbo Yu et al., "Fully Moderated T-Statistic for Small Sample Size Gene Expression Arrays," Statistical Applications in Genetics and