Biology 644

Old Title: Bioinformatics for Molecular Biologists

Potential New Title: Integrated Bioinformatics Using R for Both Wet and Dry Scientists

Outline

- Make-up Classes
- Problem Set 3 and 4
- Simpler (Univariate) Regression
- Multivariate Regression
- Generalized Models
- One-way analysis of variance
- R Lab
  - Chapter 5
  - Chapter 5 Supplemental

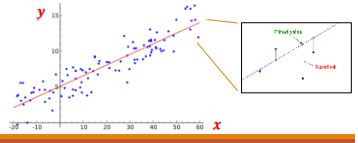
Simple (univariate) Linear Regression

- simple linear regression fits a straight line through the set of n points in such a way that makes the sum of squared residuals of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible.
- Suppose there are n data points $\{y_i, x_i\}$, where $i = 1, 2, \ldots, n$. The goal is to find the equation of the straight line $y_i = \alpha + \beta x_i + \epsilon_i$ which would provide a "best" fit for the data points.
- The "best" will be understood as in the least-squares approach: the line that minimizes the sum of squared residuals of the linear regression model.

$$\text{Find } \min_{\alpha,\beta} Q(\alpha, \beta), \text{ where } Q(\alpha, \beta) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

- α is the intercept (optional)
- β is the slope

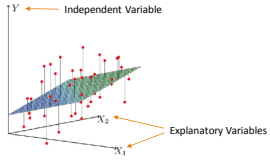
Linear least squares fitting – we seek the linear function of x that minimizes the sum of squared residuals from y.



Multiple Linear Regression

- Modeling of the linear relationship between a scalar dependent variable y and a set of more than one explanatory (independent) variables denoted as X.
- Now the set of data points is $\{y_i, x_{i1}, \ldots \ldots, x_{ip}\}$, where i = 1, 2, …, n, j = 1,2, …, p. The goal is to find the equation of the hyperplane $y_i = \alpha + \beta_1 x_{i1} + \ldots \ldots + \beta_p x_{ip} + \epsilon_i$ that provides a "best" fit for the data points.
- The "best" will be understood as in the least-squares approach: the hyperplane that minimizes the sum of squared residuals of the linear regression model.
- α is the intercept (optional)
- β_j is the coefficient (slope) for x_j

Linear least squares fitting – we seek the linear function of X that minimizes the sum of squared residuals from Y.



Multiple Linear Regression

- Modeling of the linear relationship between a scalar dependent variable y and a set of more than one explanatory (independent) variables denoted as X.
- Now the set of data points is $\{y_i, x_{i1}, \ldots \ldots, x_{ip}\}$, where i = 1, 2, …, n, j = 1,2, …, p. The goal is to find the equation of the hyperplane $y_i = \alpha + \beta_1 x_{i1} + \ldots \ldots + \beta_p x_{ip} + \epsilon_i$ which would provide a "best" fit for the data points.
- The "best" will be understood as in the least-squares approach: the hyperplane that minimizes the sum of squared residuals of the linear regression model.
- α is the intercept
- β_j is the coefficient (slope) for x_j
- when $^{\mathsf{T}}$ denotes the transpose, so that $x_i^{\mathsf{T}}\boldsymbol{\beta}$ is the inner product between vectors x_i and $\boldsymbol{\beta}$ we get:

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

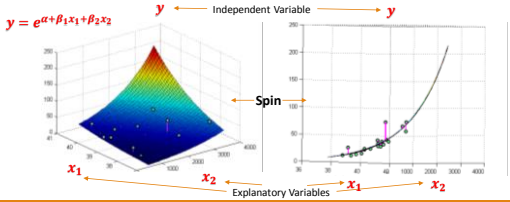
Often these *n* equations are stacked together and written in vector form

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

Generalized Linear Model

- Modeling of the linear relationship between a scalar dependent variable y and a set of more than one explanatory (independent) variables denoted as X.
- Now the set of data points is $\{y_i, x_{i1}, \ldots \ldots, x_{ip}\}$, where i = 1, 2, …, n, j = 1,2, …, p. The goal is to find the equation of the hypersurface $\mathbb{E}(\eta(y_i)) = \alpha + \beta_1 x_{i1} + \ldots \ldots + \beta_p x_{ip} + \epsilon_i$ that provides a "best" fit for the data points. η is the link function that's a member of the exponential function family.
- The "best" will be understood as in the MLE approach: the hypersurface that minimizes the sum of squared residuals of the regression model.



One-way analysis of variance

- The one-way analysis of variance (ANOVA) tests the null hypothesis that samples in two or more groups are drawn from populations with the same mean values. Typically, a sample is partitioned into 2 or more samples by a categorical variable.
- Tests the null hypothesis is H_0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_n$ by comparing the variability between groups to the variability within groups
- A generalization of the pooled-variance two-sample t-Test, and has the same assumptions!
- Independent simple random samples
- the populations are normally distributed
- the pooled population variances $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = ... = \sigma_n^2$ are equal
- Produces an F-statistic, the ratio of the sample variance calculated among the means to the sample variance within the samples. If the group means are drawn from populations with the same mean values, then the sample variance between the group means should be lower than the sample variance within the samples. A higher ratio therefore implies that the samples were drawn from populations with different mean values.
- When there are only two means to compare, the t-test and the one-way ANOVA are equivalent, and the relation between the F-test and the t –test is given by $F = t^2$.