Scientific Computing for Biologists

Yet More ANOVA and Regression

Instructor: Paul M. Magwene

23 September 2014

Overview of Lecture

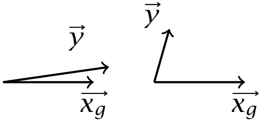
- More Complex ANOVA Models as Projection Operations
- Logistic Regression
- LOESS Regression

Hands-on Session

- ANOVA
- Logistic Regression
- LOESS

Reminder: Two-group One-way ANOVA as Regression

- Setup a ‘dummy variable’ as the predictor X_g. We assign all subjects in group 1 the value 1 and all subjects in group 2 the value -1 on the dummy variable. We then regress the variable of interest, Y, on X_g.
- When the means are different in the two groups, X_g will be a good predictor of the variable of interest, hence $\vec{y}$ and $\overrightarrow{x_g}$ will have a small angle between them.
- When the means in the two groups are similar, the dummy variable will *not* be a good predictor. Hence the angle between $\vec{y}$ and $\overrightarrow{x_g}$ will be large.



Different means

Similar means

Multi-group One-way ANOVA as Regression

- Exactly the same idea applies to g groups, except now instead of one grouping variable, we define $g-1$ grouping variables, $\dim(X_g) = g-1$.
- Then we calculate the multiple regression as we did before:

$$\mathsf{y} = \mathsf{X}\mathsf{b} + \mathsf{e}$$

$$\mathsf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} ; \mathsf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1g} \\ 1 & x_{21} & x_{22} & \cdots & x_{2g} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{ng} \end{bmatrix} ;$$

Estimate b as:

$$\mathsf{b} = (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\mathsf{y}$$

How Do We Construct the Grouping Matrix, X_g?

Two common methods are:

1. Dummy coding – define a set of g grouping variables, where values take either 0 or 1, depending on group membership, but *use only the first* $g-1$ columns:

$$U_j = \begin{cases} 1, & \text{for every subject in group } j, \\ 0, & \text{for all other subjects.} \end{cases}$$

   and

$$X_g = [U_1, U_2, \cdots, U_{g-1}]$$

2. Effect (deviation) coding – define the U_j as above, and set:

$$X_g = [U_1 - U_g, U_2 - U_g, \cdots, U_{g-1} - U_g]$$

In general, effect coding is more similar to standard ANOVA contrasts. See this ☞ web-page for a more in depth discussion of different coding schemes.

ANOVA: Example Data Set

	g_1	g_2	g_3	g_4	
	20	21	17	8	
	17	16	16	11	
	17	14	15	8	
$M_{g.}$	18	17	16	9	$M_{..} = 15$

$$y = \begin{bmatrix} 20 \\ 17 \\ 17 \\ 21 \\ 16 \\ 14 \\ 17 \\ 16 \\ 15 \\ 8 \\ 11 \\ 8 \end{bmatrix}, \qquad \mathrm{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix}$$

ANOVA: Example Data Set, cont

Solving for b we find:

$$\mathrm{b} = \begin{bmatrix} 15 \\ 3 \\ 2 \\ 1 \end{bmatrix}, \quad |\hat{y}|^2 = 150, \ |e|^2 = 40$$

Since, $\dim(\mathcal{V}_x) = 3$, and $\dim(\mathcal{V}_e) = 8$, we get:

$$F = \frac{\dim(\mathcal{V}_e)|\vec{\hat{y}}|^2}{\dim(\mathcal{V}_x)|\vec{e}|^2} = 10$$

Here's the more conventional ANOVA table for the same data:

Source	df	SS	MS	F	Pr(F)
Experimental	3	150	50	10	.0044
Error	8	40	5		
Total	11	190			

More Complex ANOVA models

- Multi-way ANOVA – used when samples are classified with respect to two or more factors (grouping variables). Allow for exploring interactions between facdtors.
- Nested ANOVA – used when there is more than one grouping variable, and the grouping variables form a nested hieararchy (groups, subgroups, subsubgroups)

As before, all of these can be treated as regression problems with appropriate design matrices!

Logistic Regression

Logistic regression is used when the dependent variable is discrete (often binary). The explanatory variables may be either continuous or discrete.

Examples:

- whether a gene is turned off (=0) or on (=1) as a function of levels of various proteins
- whether an individual is healthy (=0) or diseased (=1) as a function of various risk factors.

Model the binary responses as:

$$P(Y = 1 | X_1, \ldots, X_p) = g^{-1}(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$

So we're modeling the probability of the states as a function of a linear combination of the predictor variables.

Logistic Regression, Logit Transform

Most common choice for g is the ‘logit link’ function:

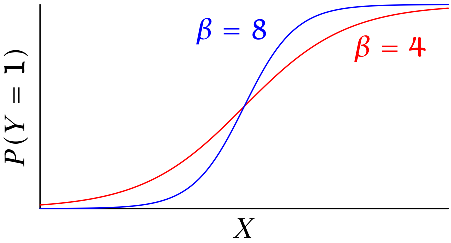
$$g(\pi) = log\left(\frac{\pi}{1-\pi}\right)$$

and g^{-1} is thus the logistic function:

$$g^{-1}(z) = \frac{e^{z}}{1+e^{z}}$$

Logistic Regression

$$P(Y = 1|X) = \frac{e^{X\beta}}{1 + e^{X\beta}}$$



Notes on Logistic Regression

- The regression is no longer linear
- Estimating the β in logistic regression is done via maximum likelihood estimation (MLE)
- Information-theoretic metrics of model fit rather than F-statistics

Logistic Regression Example

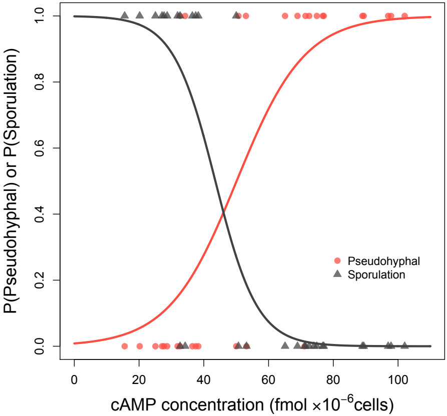
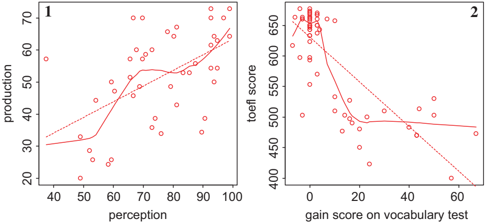


Figure: Logistic regression for yeast developmental phenotypes as a function of cAMP concentration.

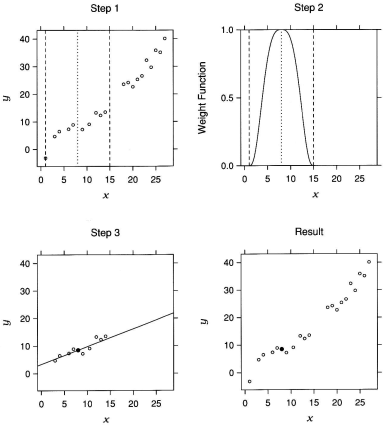
Loess Regression

- A type of non-parametric regression
- Basic idea – fit a curve (or surface) to a set of data by fitting a large number of *local regressions*.
- Cleveland, W.S. (1979). “Robust Locally Weighted Regression and Smoothing Scatterplots". Journal of the American Statistical Association 74 (368): 829-836. doi:10.2307/2286407.



Graphical overview of Loess fitting, I

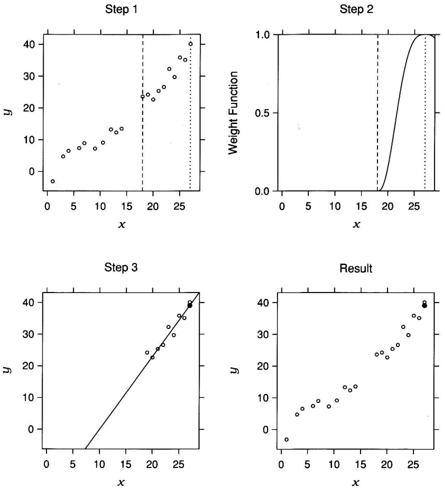
from Cleveland (1993)



3.49 HOW LOESS WORKS. The graphs show how the initial fit at $x = 8$ is computed. (Top left) α, which is 0.5, is multiplied by 20, the number of points, which gives 10. A vertical strip is defined around $x = 8$ so that one boundary is at the 10th nearest neighbor. (Top right) Weights are defined for the points using the weight function. (Bottom left) A line is fitted using weighted least-squares. The value of the line at $x = 8$ is the initial loess fit at $x = 8$. (Bottom right) The result is one point of the initial loess curve, shown by the filled circle.

Graphical overview of Loess fitting, II

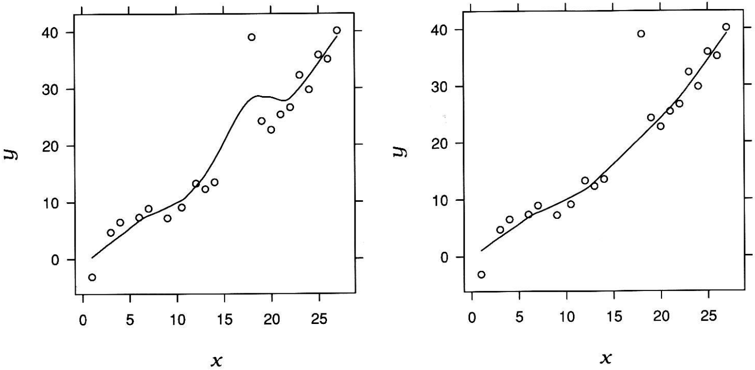
from Cleveland (1993)



3.50 HOW LOESS WORKS. The computation of the initial loess fit value at $x = 27$ is illustrated.

Graphical overview of Loess fitting, III

from Cleveland (1993)



3.51 HOW LOESS WORKS. Loess employs robustness iterations that prevent outliers from distorting the fit. (Left panel) The open circles are the points of the graph; there is one outlier between x = 15 and x = 20. The initial loess curve has been distorted in the neighborhood of the outlier. (Right panel) The graphed curve is the fit after four robustness iterations. Now the fit follows the general pattern of the data.