Scientific Computing for Biologists

Linear Algebra Review II & Regression

Instructor: Paul M. Magwene

20 September 2011

Overview of Lecture

- More Linear Algebra
  - Linear combinations and Spanning Spaces
  - Subspaces
  - Basis vectors
  - Dimension
  - Rank
- More on Regression
  - Multiple regression
  - Curvilinear regression
  - Logistic regression
  - Major axis regression

Hands-on Session

- Regression in R
- Multiple regression
- Logistic regression
- Locally weighted regression (LOESS or LOWESS)

Space Spanned by a List of Vectors

Definition

Let X be a finite list of n-vectors. The **space spanned** by X is the set of all vectors that can be written as linear combinations of the vectors in X.

A space spanned includes the zero vector and is closed under addition and multiplication by a scalar.

Remember that a *linear combination* of vectors is an equation of the form $z = b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \cdots + b_p\mathbf{x}_p$

Subspaces

$\mathbb{R}^n$ denotes the seat of real n-vectors - the set of all $n \times 1$ matrices with entries from the set $\mathbb{R}$ of real numbers.

Definition

A **subspace** of $\mathbb{R}^n$ is a subset S of $\mathbb{R}^n$ with the following properties:

1. $\mathbf{0} \in S$
2. If $\mathbf{u} \in S$ then $k\mathbf{u} \in S$ for all real numbers k
3. If $\mathbf{u} \in S$ and $\mathbf{v} \in S$ then $\mathbf{u} + \mathbf{v} \in S$

Examples of subspaces of $\mathbb{R}^n$:

- any space spanned by a list of vectors in $\mathbb{R}^n$
- the set of all solution to an equation $A\mathbf{x} = \mathbf{0}$ where A is a $p \times n$ matrix, for any number p.

Basis

Let S be a subspace of $\mathbb{R}^n$. Then there is a finite list, X of vectors from S such that S is the space spanned by X.

Let S be a subspace of $\mathbb{R}^n$ spanned by the list $(u_1, u_2, \ldots, u_n)$. Then there is a linearly independent sublist of $(u_1, u_2, \ldots, u_n)$ that also spans S.

Definition

A list X is a **basis** for S if:

- X is linearly independent
- S is the subspace spanned by X

Dimension

Let S be a subspace of $\mathbb{R}^n$.

Definition

The **dimension** of S is the number of elements in a basis for S.

Rank of a Matrix

Let A by an $n \times p$ matrix.

Definition

The **rank** of A is equal to the dimension of the row space of A which is equal to the dimension of the column space of A.

Where the row space of A is the space spanned by the list of rows of A and the column space of A is defined similarly.

Equivalence Theorem

Let A by an $p \times p$ matrix. The following are equivalent

- A is singular
- the rank of A is less than p
- the columns of A form a LD list in $\mathbb{R}^n$.
- the rows of A form a LD list in $\mathbb{R}^n$
- the equation $A\mathbf{x} = \mathbf{0}$ has non-trivial solutions
- the determinant of A is zero

Regression Models

Variable space view of multiple regression

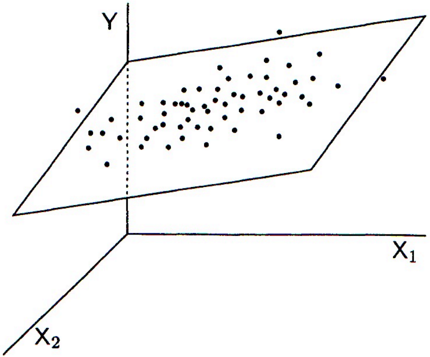
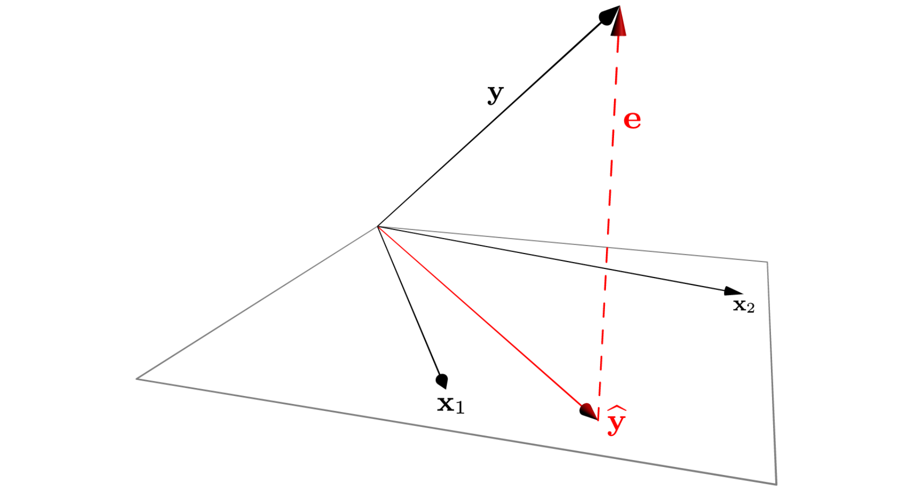


Figure 4.1: *The regression of* Y *onto* X_1 *and* X_2 *as a scatterplot in variable space.*

Subject Space Geometry of Multiple Regression



Multiple Regression

Let Y be a vector of values for the outcome variable. Let $\mathbf{X}_i$ be explanatory variables and let $\mathbf{x}_i$ be the mean-centered explanatory variables.

$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e}$$

where –

Uncentered version:

$$\hat{Y} = a\mathbf{1} + b_1\mathbf{X}_1 + b_2\mathbf{X}_2 + \cdots + b_p\mathbf{X}_p$$

Centered version:

$$\hat{y} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \cdots + b_p\mathbf{x}_p$$

Statistical Model for Multiple Regression

In matrix form:

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

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

$$\mathbf{b} = \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \; ; \; \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Estimating the Coefficients for Multiple Regression

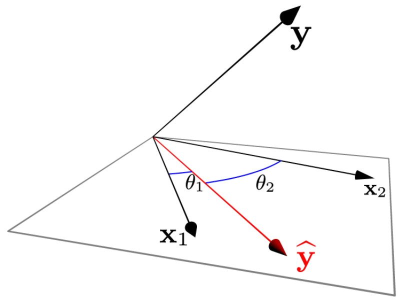
$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

Estimate **b** as:

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

Multiple Regression Loadings

The regression **loadings** should be examined as well as the regression coefficients.



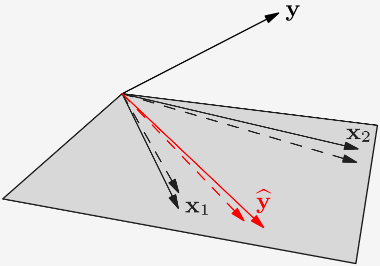
Loadings are given by:

$$\cos\theta_{\vec{x_j},\vec{\widehat{y}}} = \frac{\vec{x_j}\cdot\vec{\widehat{y}}}{|\vec{x_j}||\vec{\widehat{y}}|}$$

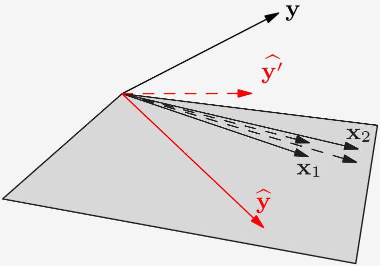
Multiple regression: Cautions and Tips

- Comparing the size of regression coefficients only makes sense if all the predictor variables have the same scale
- The predictor variables (columns of $\mathbf{X}$) must be linearly independent; when they're not the variables are **multicollinear**
- Predictor variables that are **nearly multicollinear** are, perhaps, even more difficult to deal with

Why is near multicollinearity of the predictors a problem?



(a) Non-collinear predictors



(b) Nearly collinear predictors

Figure: When predictors are nearly collinear, small differences in the vectors can result in large differences in the estimated regression.

What can I do if my predictors are (nearly) collinear?

- Drop some of the linearly dependent sets of predictors.
- Replace the linearly dependent predictors with a combined variable.
- Define orthogonal predictors, via linear combinations of the original variables (PC regression approach)
- 'Tweak' the predictor variables so that they're no longer multicollinear (Ridge regression).

Curvilinear Regression

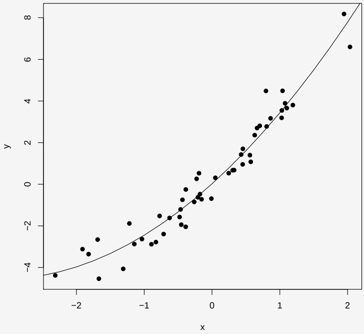
Curvilinear regression using **polynomial models** is simply multiple regression with the x_i replace by powers of x.

$$\hat{y} = b_1\mathbf{x} + b_2\mathbf{x}^2 + \cdots + b_p\mathbf{x}^n$$

Note:

- this is still a *linear* regression (linear in the coefficients)
- best applied when a specific hypothesis justifies there use
- generally not higher than quadratic or cubic

Example of Curvilinear Regression



$$\mathbf{y} = 3\mathbf{x} + 0.5\mathbf{x}^2 + \mathbf{e}$$

```
lm(formula = y ~ x + I(x^2))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02229    0.11651   0.191    0.849
x            2.94001    0.09693  30.331  < 2e-16 ***
I(x^2)       0.47146    0.07685   6.135 1.68e-07 ***
```

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:

$$P(Y = 1 | X_1, \ldots, X_p) = g^{-1}(\beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \cdots + \beta_p \mathbf{x}_p)$$

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:

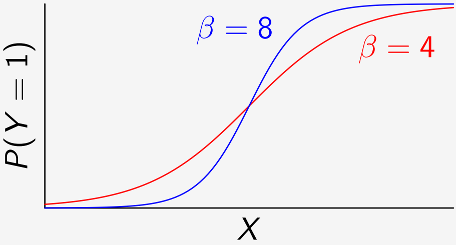
$$g(\pi) = log\left(\frac{\pi}{1-\pi}\right)$$

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

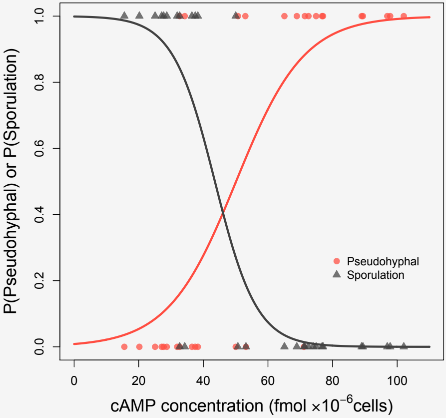
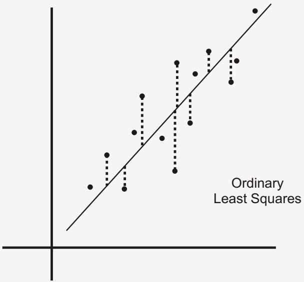
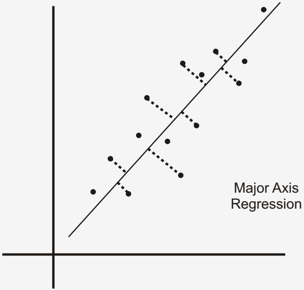


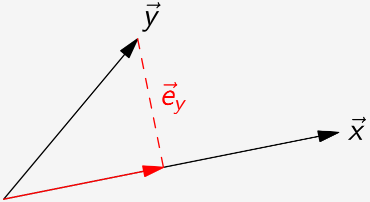
Figure: Logistic regression for yeast developmental phenotypes as a function of cAMP concentration.

OLS vs. Major Axis Regression

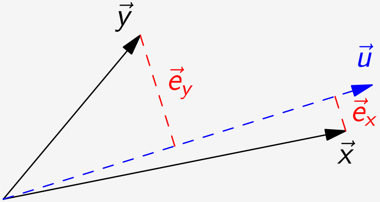




Vector Geometry of Major Axis Regression



(a) OLS



(b) Major Axis Regression

Figure: Vector geometry of ordinary least-squares and major axis regression.