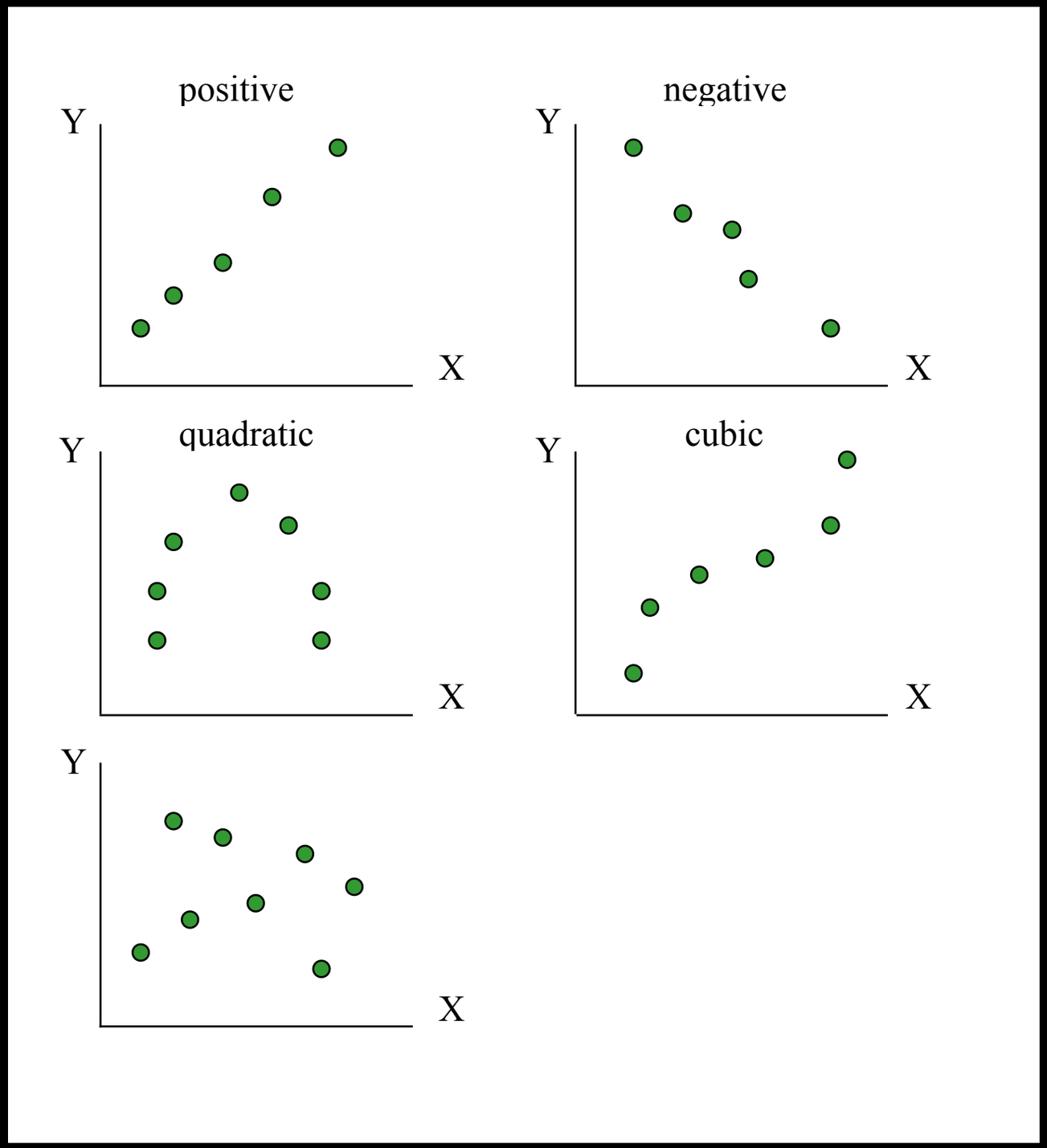
Continuous data and linear models

Katie Pollard

BMI 206
September 26, 2016

Relating Continuous Variables

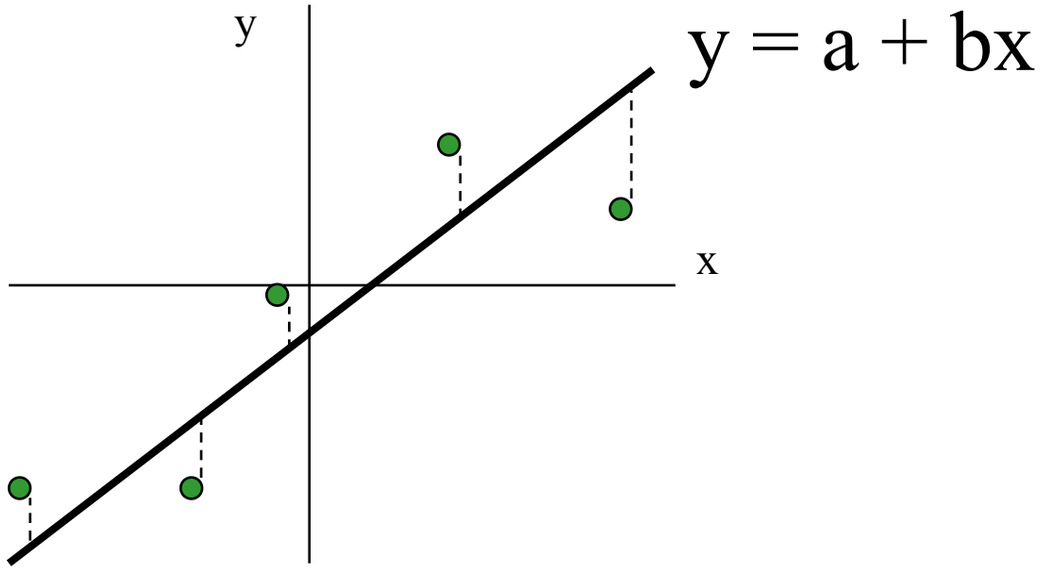


Linear relationship

Non-linear relationship

No relationship

Linear Model



a is the intercept
b is the slope

Seek the line that minimizes sum of squared residuals.

The solution to the least squares problem is:

$$a = \bar{y} - b\bar{x} \qquad b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

- Substituting estimates of (a,b) provides predictions.
- Residual is observed minus predicted value for each x.

Data Transformations

If Y increases a non-constant amount per unit increase in X, transformation may produce a linear relationship:

- Log or exponentiate
- Root or raise to a power
- Reciprocal
- Z-scores (subtract mean, divide by standard error)

For non-continuous data (e.g., counts), other models are typically needed. Generalized linear models will be covered next week.

Correlation

Pearson's correlation coefficient (ρ) is estimated by:

$$r = \frac{\sum_{i=1}^{n} z_x(i) z_y(i)}{n-1}$$

- Quantifies <u>linear</u> X vs. Y relationship.
- -1≤ r ≤ 1
- Positive (r>0) if positive slope
- Negative (r<0) if negative slope
- r=0 if no <u>linear</u> relationship (may have other relationship)
- Coefficients in linear models measure correlation

Spearman's correlation and Kendall's tau are more robust. They measure rank correlation (monotonicity).

Multiple Regression

- One outcome variable, 2+ covariates
  - Covariates can be continuous or categorical
  - May include powers or other transformations of covariates
  - May include interactions between covariates
- Coefficients represent expected change in Y per unit increase in that covariate, while holding the other covariates constant (i.e., adjusted for them).
- X and Y can be conditionally associated, but marginally independent. Or the opposite.
- Coefficient estimates and their standard errors can be used to test for association with Y.
- Predicted values can be computed for different covariate combinations.

Evaluating Model Fit

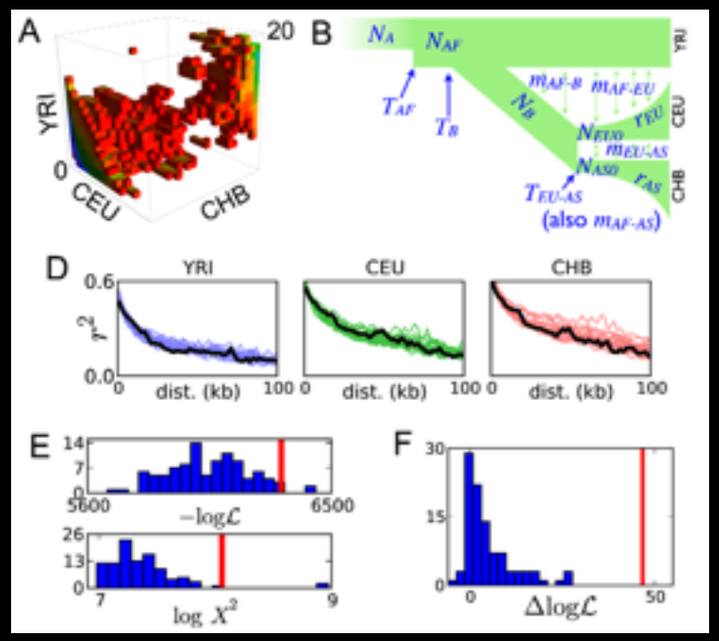
- Plot of residuals vs. X: no trends if linear relationship
- Influential points: big effect on estimates, usually small residual
- Model diagnostics quantify fit:

$$r^2 = \frac{SS_{total} - SS_{resid}}{SS_{total}} = 1 - \frac{SS_{resid}}{SS_{total}} \qquad s_e = \sqrt{\frac{SS_{resid}}{n-2}}$$

- Coefficient of determination (r^2) is the amount of variation in Y that cannot be explained by the linear relationship (i.e., model) between X and Y.
- Pearson's correlation coefficient (r) is the square root of the coefficient of determination.
- Standard deviation about the least squares line (s_e) is average distance points are from the line.

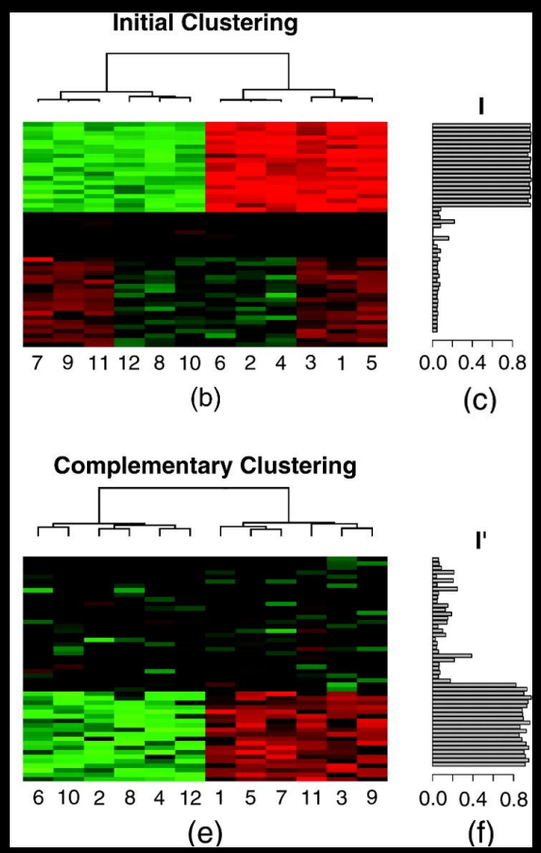
Code Examples

Model selection in bioinformatics



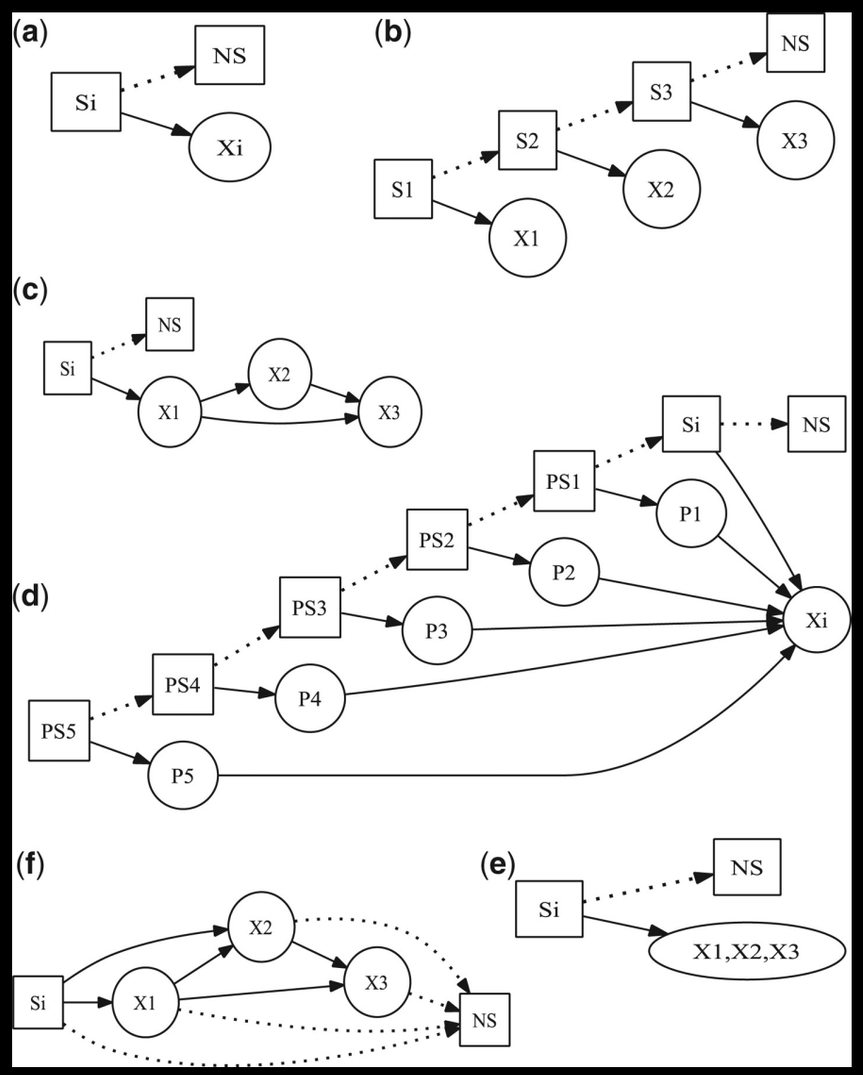
Gutenkunst et al. (2009) PLoS Genetics

A model of human expansion out of Africa that includes contemporary migration fits polymorphism data from Africa, Europe, and China better than a model without migration.



Nowak & Tibshirani (2007) Biostatistics

Clustering results and variable importance are very different after down-weighting highly expressed genes. The groups identified have different associations with survival and prognostic variables.



Mork & Holmes (2012) Bioinformatics

A collection of HMMs for modeling bacterial protein-coding gene potential.

Model selection in bioinformatics

In bioinformatics, typically many covariates are measured:

- Expression of thousands of genes in each tumor
- Genotypes at millions of variants in each patient
- Evolutionary signatures at hundreds of pairs of residues in each protein structure
- Abundance of hundreds of thousands of proteins in each metagenome

Should all of these variables be in a model for an outcome of interest? Can they even feasibly be included?

With so many variables, over-fitting the observed data is a serious risk. This reduces generalizability of the results.

Model Selection Criteria

The goal of model selection is to pick the best model given the data.

- Different criteria for evaluating “best”
  - Likelihood ratio statistic (compare to chi-square for test)
  - Change in residual sum of squares
  - R^2 or adjusted R^2
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)

- The last three account for the number of parameters with penalties to avoid over-fitting or too complex models.

Additional Validation

Even with penalties for large models, observed data can be overfit, reducing generalizability and repeatability of results.

Some solutions:

- Cross-validation involves holding out a random subset of the data and assessing model fit on the held out data, repeatedly.

- External validation involves assessing model fit on a totally independent data set (e.g., from a replication study or another population)

For both, can also use prediction accuracy as criterion.

Algorithms for searching large space of possible models

All subsets selection involves enumerating all possible models and picking the best one. Some times this is computationally infeasible. Alternatives include

- Forward selection: Start with a small model and build up
- Backward selection: Start with the full model and remove terms
- Forward-backward selection: After building up, try removing terms to see if fit improves
- Deletion-substitution-addition: Algorithm for searching in a less linear fashion

Additional issue: Include interactions without main effects?

Correlated variables

In bioinformatics, covariates are often highly correlated:

- Co-expressed genes
- Genotypes in haplotype blocks
- Evolutionary signatures at adjacent residues
- Proteins in the same pathway or macromolecule

Consequently, only one (or a few) correlated variables will typically be selected for the model.

What determines which variable is selected?

Is the selected variable more important biologically?

Variable importance

The importance of each covariate towards model fit can be measured with various statistics, e.g.:

- Estimated coefficient divided by its standard error (linear models - this fails in many other models)
- Sign of coefficient
- Fit of model with and without the variable included
- Average error minus error after permuting the covariate values, divided by its standard error (in cross-validation)
- Decrease in node impurity after adding variable (random forests)

For classification, assess importance for each class.