Statistical Practices and Research for Interdisciplinary Sciences (**SPRIS**)

Lecture 4

Yuanjia Wang, Ph.D.

Department of Biostatistics, Mailman School of Public Health
Columbia University

& Division of Biostatistics, New York State Psychiatric Institute

Identify the objective and whether the data are appropriate for the proposed analyses.

Before any formal analysis:

- Understand the design (e.g., observational study versus experiment)
- Understand the objective (e.g., hypothesis testing, estimation of association, prediction of future independent outcomes): inference (parameter estimation consistency [bias] and variance, coverage of confidence intervals) versus prediction (prediction error, RMSE, accuracy, AUC).
- Decide if regression analysis is appropriate

Stratification, Matching, Regression

Consider an observational study comparing patients receiving treatments A and B. Females are more likely to receive treatment B, so need to adjust for sex. Methods for adjustment:

- Regression approach: fit a model with covariates including treatment and sex; treatment effect is adjusted for sex
- Stratification approach: for males, estimate the treatment group difference and also for females; average of the two differences is adjusted for sex
  - For a continuous variable to be adjusted (e.g., age), can consider discretizing based on quintiles and treated as a categorical variable.
  - Stratification is often used in randomized trials to ensure that randomization stays balanced within subsets of subjects
- Matching

Stratification, Matching, Regression

Matching approach: match a patient receiving A similar to a patient receiving B (e.g., within x years of age). Perform a matched pairs analysis (e.g., paired t-test).

- Matching may also be a useful design (e.g., matched case-control design for rare outcomes), but analysis may lead to reduced sample size and power.

Matching and stratification are not efficient in the presence of many adjustment variables

- Consider matching or stratification by propensity scores (probability of receiving treatment A given covariates)

covariance balance may be bad for some covariates

Regression analysis may have advantages with many covariates. It requires more assumptions, but

"All models are wrong, but some are useful". –George Box

When is Regression Used?

- Hypothesis testing: Test for no association (correlation) of a predictor and a response variable (unadjusted test) or test for no association of predictor and response after adjusting for the effects of other predictors
- Estimation: estimate the conditional mean of the outcome given covariates
- Prediction:
  - Predicting central tendencies given covariates, e.g., long-term average response as a function of predictors
  - Predicting individual outcomes
- Increasing power and precision for assessing the effect of a target variable by including other prognostic factors that correlate with the outcome variable. variance of beta depends on the residuals
- Confounder adjustment: getting adjusted estimates of effects

Regression Analysis Tasks

- Preliminary analysis (e.g., plots, transformation)
- Model selection
- Estimation and refinement of the model
- Model diagnostics and evaluation

Preliminary steps

- Plot the data (e.g., bivariate associations)
- Are all the predictors necessary? Are all confounders included?
- How are missing data coded, and how should they be treated?
- How should outliers be treated?
- Are any of the categorical variables properly coded?
- Are observations independent?

Considerations in a Regression Analysis

- Plots: Scatter plots of Y versus each of the X_i. Pairwise plots (X_i versus X_k.
- Number of Predictors: Overfitting a regression model has the adverse effect of inflating the variance associated with the parameter estimates. Multicollinearity problems may also arise due to linear dependencies between the predictors. Justifying the need for each predictor at the outset is valuable.
- Missing Values: Predictors containing many missing values may be dropped from the regression model or consider multiple imputations.
- Outliers: Detect influential points and consider remove
- Recognize repeated measures, longitudinal data, or clustering of observations
- Purpose of regression: achieving accurate prediction at the individual level may require a larger sample size (e.g., deep learning models)

Fitting Regression Models

Regression analysis is an iterative process, alternating between assessing the quality of the fitted model and making modifications to the model. However, it is worth emphasizing:

Don't overdo the fitting procedure.

Running the risk of overfitting. An unlimited number of regression models and many diagnostic checks can be performed. In practice, the best model can be the one that has the simplest interpretation.

Types of Regressions

Classical linear regression

Nonlinear model, nonparametric regression

Machine learning methods: tree-based methods, random forests, deep learning

The Classical Linear Model

The model: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
Assumptions:

1. Covariates X are treated as fixed, expectation of errors: $E(\boldsymbol{\varepsilon}) = \mathbf{0}$.
2. Variances and correlation structure of errors

$$Cov(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\boldsymbol{I},$$

   that is homoscedasticity and uncorrelated errors.
3. The design matrix $\boldsymbol{X}$ has full column rank.
4. Independent, additive Gaussian errors: $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$.

When all the covariates are random (i.e., for observational studies instead of designed experiments), all assumptions are conditional on $\boldsymbol{X}$, i.e.,

$$E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \mathbf{0}, E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}) = \sigma^2\boldsymbol{I},$$

and X_i and ε_i are independent.

Discussion of Assumptions

1. The errors have (conditional) expectation zero, implying the systematic trend is captured in a linear function of $\boldsymbol{X}$ (i.e., $E(Y|\boldsymbol{X}) = \boldsymbol{\beta}^T\boldsymbol{X}$). If violated, consider the transformation of covariates or nonlinear/nonparametric regression.

2. The homescedasticity assumption implies error variance does not vary across individuals. Can be relaxed by assuming:

$$Cov(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\boldsymbol{W},$$

where $\boldsymbol{W} = \text{diag}(w_1^2, \cdots, w_n^2)$. Consider how to parametrize w_j^2 and ensure it can be estimated from data.

3. Full rank condition ensures identifiability (no collinearity). Classical regression: $p < n$

4. Uncorrelated assumptions can be violated in the presence of autocorrelation. Most often encountered in time series or longitudinal data.

Autocorrelation

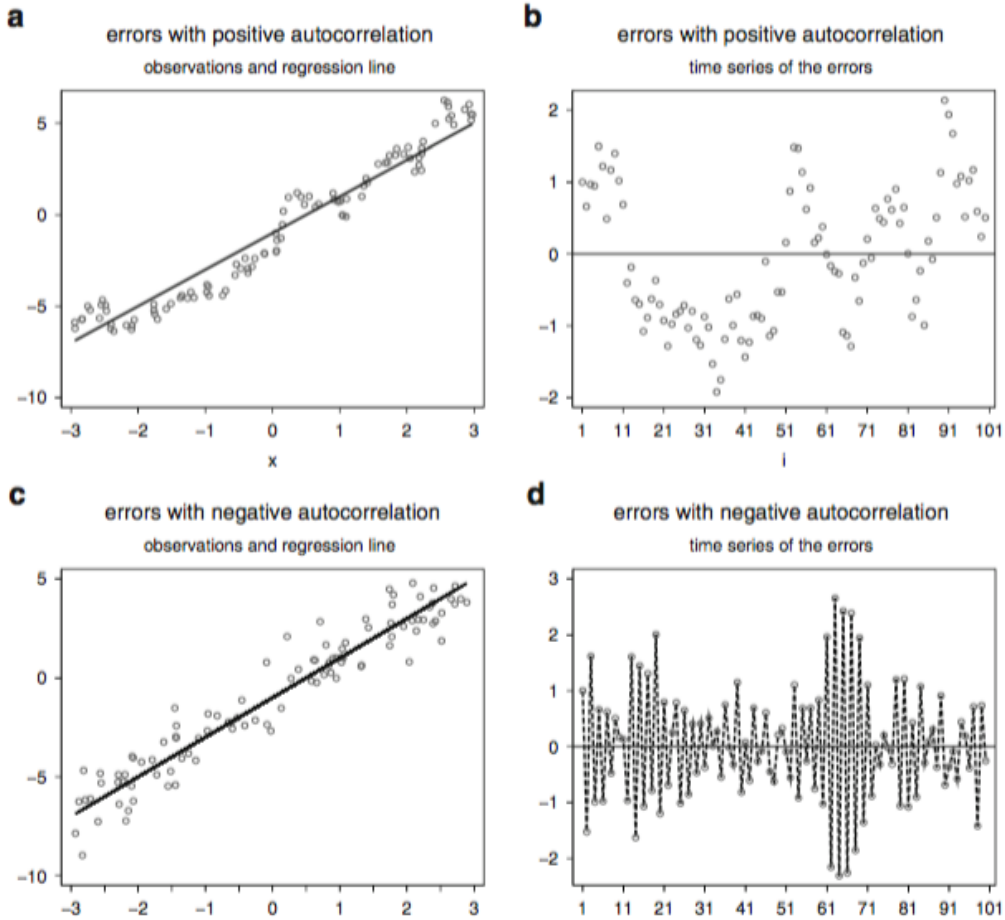


Fig. Positively correlated errors (top): $y_i = -1 + 2x_i + \varepsilon_i, \varepsilon_i = 0.9\varepsilon_{i-1} + u_i$.
Negatively correlated errors (bottom): $\varepsilon_i = -0.9\varepsilon_{i-1} + u_i$.

Autocorrelation

Autocorrelated errors can appear when the regression model is misspecified, e.g., a covariate is missing, or the effect of a continuous covariate is nonlinear.

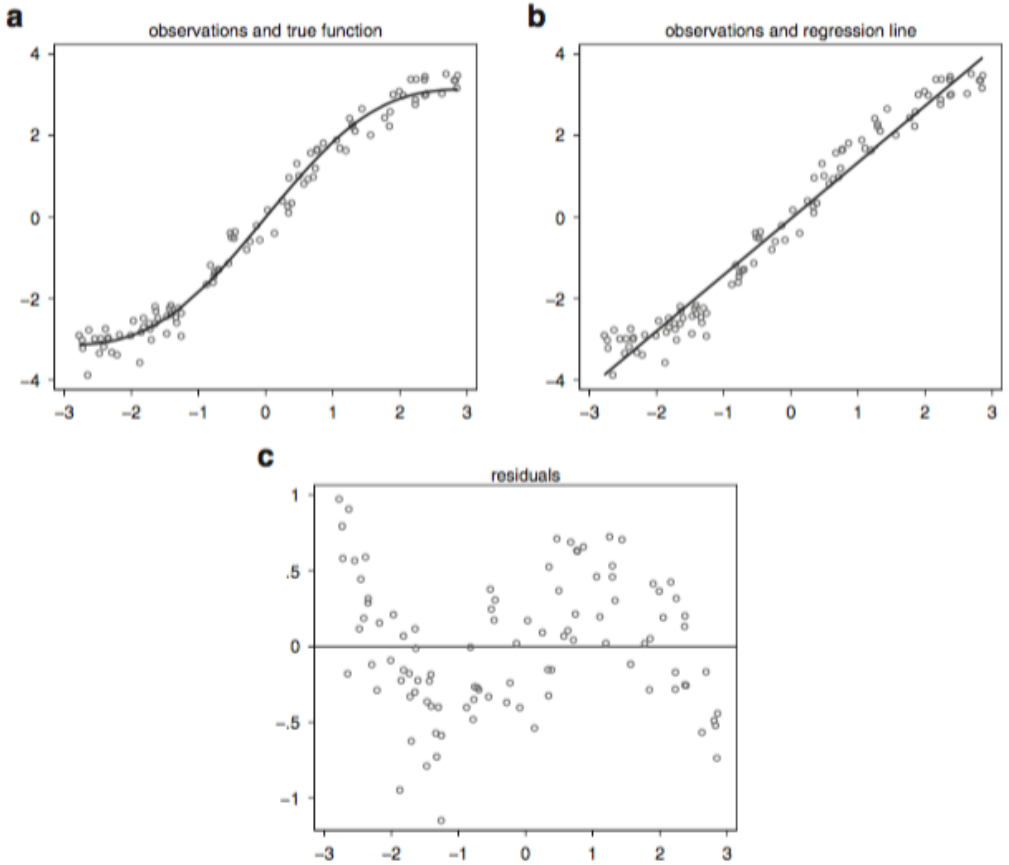
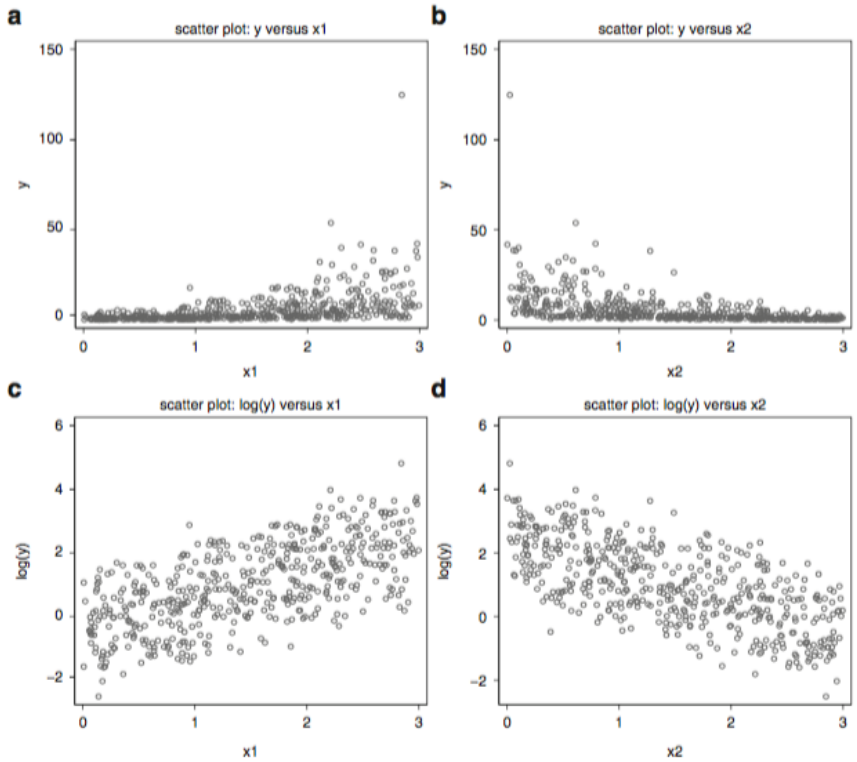


Fig. Correlated residuals when the model is misspecified: $E(Y_i|X_i) = \sin(X_i) + X_i + \varepsilon_i$. (b) shows the estimated regression line, i.e., the nonlinear relationship is ignored.

Additive Error Assumption

In theory, many different models for the structure of the errors are conceivable. In most applications, two alternative error structures are assumed: additive errors and multiplicative errors (log-normal).

Multiplicative error model: $Y_i = \exp(\boldsymbol{\beta}^T X_i + \varepsilon_i) = \exp(\boldsymbol{\beta}^T X_i)\exp(\varepsilon_i)$, $\log(Y_i) = \boldsymbol{\beta}^T X_i + \varepsilon_i$



The Classical Linear Model

Estimation: least squares $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ or weighted least squares.

Geometric interpretation: minimizes distance from Y to model space spanned by X (projection).

Properties of least squares estimator:

- Least squares is also the maximum likelihood estimator of $\boldsymbol{\beta}$ when normality holds.
- Gauss-Markov theorem: LS is the best linear unbiased estimator with minimal variance (BLUE) among all linear and unbiased estimators.

Inference based on normality assumption or asymptotic theory. Use resampling techniques if necessary (permutation test, bootstrap confidence intervals)

Model diagnostics based on residual plots.

Interpretation

In a regression model,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i,$$

the interpretation of regression coefficients is based on statistical association. Causal conclusions require stronger assumptions (sometimes cannot be verified) than association.

May not be interested in interpreting coefficients for all variables (e.g., confounders).

In a controlled experiment, X_1 is treatment and other pre-treatment baseline prognostic factors that are not intervened on.

In observational studies, no control over which variables will be perturbed and which will be unperturbed. In most cases, β_k does not have causal interpretation.

Causation

Causal inference is an active research field in epidemiology and biostatistics.

Let $T = 0$ denote control and $T = 1$ denote treatment. Let Y_i^T denote the potential outcome for subject i when T applies. The causal effect for the ith patient is

$$\delta_i = Y_i^1 - Y_i^0.$$

Fundamental problem of causal inference: cannot apply both treatment and control simultaneously so do not observe both potential outcomes. Only see one of the pair (Y_i^1, Y_i^0).

Impossible to estimate individual causal effects δ_i, but in randomized controlled trials we can estimate average treatment effect (ATE), $\delta = E(\delta_i)$, over groups by randomly assign subjects to treatment or control.

Why Does Randomization Allow Estimating ATE?

Subjects will vary in ways that may affect the response (age, general health status, lifestyle). Some of these variables may not be measurable.

Randomization ensures:

- Potential outcomes are independent of treatment assignments!

$$Y_i^1, Y_i^0 \perp\!\!\!\perp T_i \tag{1}$$

- Two treatment groups are balanced in terms of both observed and unobserved confounders.

ATE can be estimated from observed RCT data under assumptions Stable Unit-Treatment-Value Assumption (SUTVA[1]) and (1):

$$\begin{aligned} E(\delta_i) &= E(Y_i^1) - E(Y_i^0) = E(Y_i^1|T_i = 1) - E(Y_i^0|T_i = 0) \\ &= E(Y_i|T_i = 1) - E(Y_i|T_i = 0) \end{aligned}$$

The second equality in red is due to (1). especially when heterogeneity exists
Caveat: Subjects are usually not selected randomly. Thus, δ may not generalize to different populations in the presence of treatment effect heterogeneity. missing data issue

[1]SUTVA: $Y_i^1 = Y_i$ when $T_i = 1$; $Y_i^0 = Y_i$ when $T_i = 0$, no interaction among units.

Observational Studies with Regression Covariate Adjustment

Subjects will vary in ways that may affect the response (age, general health status, lifestyle). If confounders are measured so that:

- potential outcomes are independent of treatment assignments given covariates: NUCA

$$Y_i^1, Y_i^0 \perp\!\!\!\perp T_i | X_i \tag{2}$$

- two treatment groups are balanced in terms of observed confounders X_i.

ATE can be estimated from observational data under assumptions SUTVA and (2):

$$\begin{aligned} E(\delta_i) &= E(Y_i^1) - E(Y_i^0) = \int E(Y_i^1 | T_i = 1, X_i) dP(X_i) - \int E(Y_i^0 | T_i = 0, X_i) dP(X_i) \\ &= \int E(Y_i | T_i = 1, X_i) dP(X_i) - \int E(Y_i | T_i = 0, X_i) dP(X_i) \end{aligned}$$

Second equality in red holds due to (2).

Observational Studies

Use matching, covariate adjustment, or inverse probability weighting (IPW) by propensity scores. Matching by propensity scores to create a 'pseudo-experiment' where measured confounders have balanced distributions between groups.

Covariate adjustment by including confounders in a regression. All approaches are invalid in the presence of unmeasured confounders. Perform sensitivity analysis.

Hill (1965)[2] summarizes qualitative support for causality:

- Strength: the magnitude of the estimated effect is large (not rely on p-value). Known covariates can be adjusted for, while unobserved and unsuspected confounding variables could easily lead to small effects. It is less likely that some unknown confounder could counteract a large effect.

[2] Hill, A. B. (1965). The environment and disease: association or causation? Proceedings of the Royal Society of Medicine 58(5), 295.

- Consistency: A similar effect has been found for different subjects under different circumstances at different times and places. Replication by independent research groups is essential in establishing causation.
- Specificity: The postulated causal factor is associated mainly with a particular response and not with many other possible responses.
- Temporality: The postulated causal factor is determined or fixed before the outcome or response is generated. Sometimes, it is unclear whether X causes Y or vice versa (reverse causality).
- Gradient: Dose-response. The response increases (or decreases) monotonely as the postulated causal variable increases.

- Plausibility: There is a credible theory suggesting a causal effect.
- Natural experiment: the researcher has not manipulated exposure to the event or intervention of interest. Used to understand the health impact of policies and other large-scale interventions. Uses difference in differences approach to estimate effect (compares change over time in exposed and unexposed groups).

Table 3 in Greenhalgh et al. 2022.

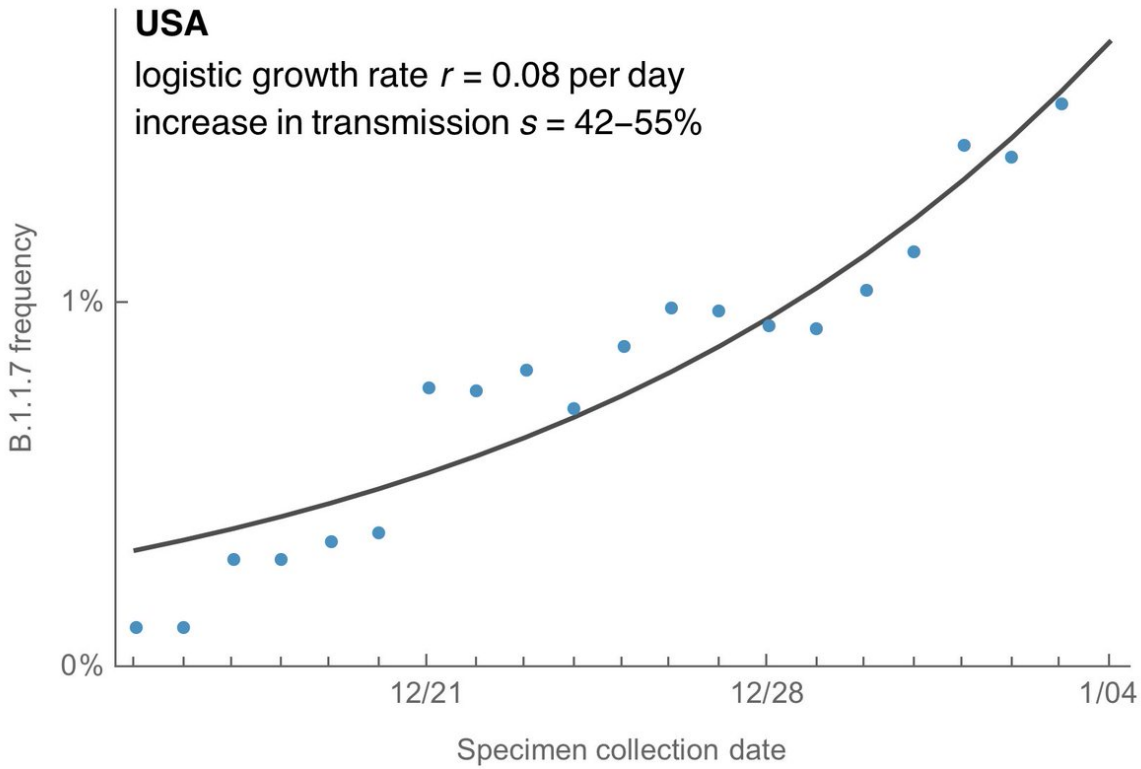
Not all of these might apply, but causation might still be present. These might apply, but we may not be sure of causation.

Nevertheless, these considerations add to the weight of the evidence.

Other Regressions: Nonlinear (Parametric) Model

Example: Modeling growth of B.1.1.7. a variant in the US by a logistic growth model (a population's growth rate gets smaller as population size approaches a maximum imposed by limited resources in the environment):

$$x(t) = \frac{1}{1 + (\frac{1}{x_0} - 1)e^{-rt}}$$



Nonparametric Regression

- Few assumptions (no assumption on the shape of the function)
- Works well with a single X Can use dimension reduction method, e.g. PCA
- Plotted trend line may be the final result of the analysis (not appropriate for prediction) only a descriptive trend
- Simplest smoother: moving average
- LOcal regrESSion (loess) superior to moving averages
- Regression splines (piecewise polynomials)

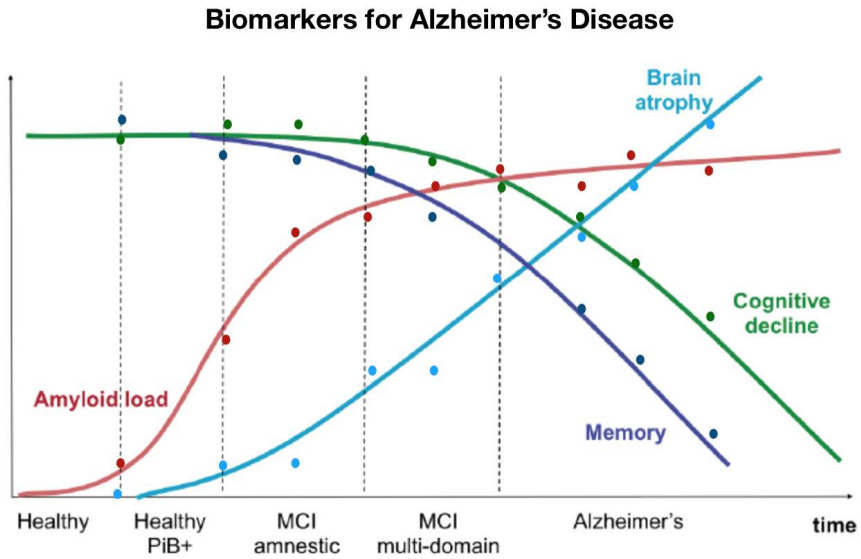


Fig. loess nonparametric smoother relating biomarkers of AD to age.

Recursive Partitioning: Tree-Based Models

Advantages:

- Nonparametric, does not require any functional form for the predictors; does not assume additivity of predictors (i.e., recursive partitioning can identify complex interactions).
- Useful in settings where overfitting is not so problematic (prediction problems), i.e., constructing propensity scores (predicting treatment assignment).

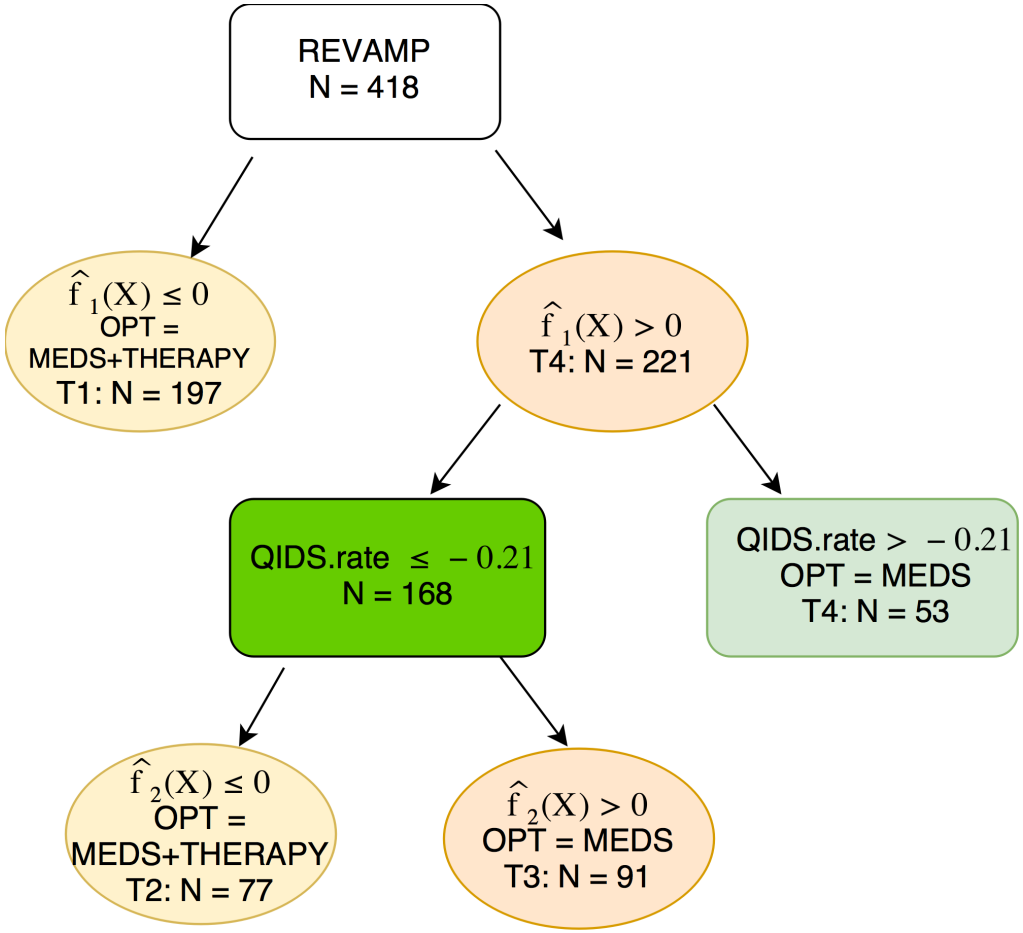
Limitations:

- Not utilizing continuous variables effectively
- Overfitting in three directions: searching for best predictors, best splits, and searching multiple times.
- High variance

Increase reliability and prediction performance: random forest.

Example: Interaction Tree For Optimizing Treatment

Composite Interaction Tree (CITree) (Qiu, X. & Wang, Y. 2019):



Prediction, Model Selection, Model Validation

Predictions

Two kinds of predictions are made from regression models: (1) predict mean response for a group of individuals; predict a future individual response.

Variability of prediction given a new subject with predictors $\boldsymbol{x}_0$:

$$\hat{y}_0 = \boldsymbol{x}_0^T \widehat{\boldsymbol{\beta}}$$

- Variance for predicting a group mean response

$$Var(\boldsymbol{x}_0^T \widehat{\boldsymbol{\beta}}) = \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0 \sigma^2$$

- Variance for predicting an individual response

$$Var(\boldsymbol{x}_0^T \widehat{\boldsymbol{\beta}}) = (1 + \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0) \sigma^2$$

Pitfalls of Inadequate Predictions

1. Poor/misspecified model (e.g., underfit, missing important predictors).
2. Quantitative extrapolation. Predict outcomes for cases with predictor values differ greatly from the data, e.g., when assessing the risk from low exposure to substances that are dangerous in high quantities (second-hand tobacco smoke, lead).
3. Qualitative extrapolation. Predict outcomes for observations that come from a different population. e.g., experimental data may not be transferred to real-world scenarios.
4. Overconfidence due to overfitting. Data analysts search around for good models for the data they have and often do too good a job in finding a fit. This can lead to unrealistically small variance estimate but model doesn't generalize.

Model Selection and Variable Selection

A large number of candidate predictors may be available. The scientific background knowledge may be insufficient to decide the model. Data-driven approaches are used.

Before performing model selection, consider following:
1. Irrelevant noise variables: What is the effect on the bias and the variance of the least squares estimator, in the case that we include noise variables in the model?

2. Missing important relevant variables: What is the effect on the bias and the variance, if we omit relevant variables in the model?

3. Prediction quality: What effect does the model specification, more specifically the selected variables in the model, have on prediction?

Effect on Least Squares Estimators

Partition $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$.

- Missing Important Variables: Even though the complete model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is correct, we mistakenly estimate the reduced model $\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}$. In this case we neglect the relevant variables $\boldsymbol{X}_2$.

Consequences of missing important variables :

- $\widehat{\boldsymbol{\beta}}_1$ is biased. An exception is the case when $\boldsymbol{X}_1^T\boldsymbol{X}_2 = \boldsymbol{0}$, i.e., every variable in $\boldsymbol{X}_1$ is uncorrelated to every variable in $\boldsymbol{X}_2$.
- The variance of estimated $\boldsymbol{\beta}_1$ based on reduced model will have a smaller variance than the corresponding LS estimator based on the correct model.
- Mean squared error (MSE) based on reduced model can be smaller, i.e., it is possible that a sparse model has better MSE than the correctly specified full model.

Effect on Least Squares Estimators

Partition $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$.

- Including noise variables: Even though the reduced model $\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}$ is correct, we mistakenly estimate the full model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. In this case, we included noise variables $\boldsymbol{X}_2$ in the model.

Consequences: Coefficient of X2 may close to zero

- Even though noise variables were considered, $\widehat{\boldsymbol{\beta}}$ is unbiased. Of course, the estimator $\widehat{\boldsymbol{\beta}}_1$ based on the true model is also unbiased.
- The estimator $\widehat{\boldsymbol{\beta}}$ has larger variance than $\widehat{\boldsymbol{\beta}}_1$. If the estimated model contains noise variables, then the precision of the estimators decreases.

Preferably, the specified model should not contain noise covariates. We might aim for a sparse model so that bias and variance, and thus MSE, are small.

Effect on Prediction

Prediction error: $E(Y_0 - \hat{Y}_{0M}) = \sigma^2 + p/n\sigma^2 + (\mu_{0M} - \mu_0)^2$.

- Irreducible Prediction Error: The first term cannot be reduced (inherent variability), even by sophisticated inference techniques.
- Variance: The second term consists of the variance of the estimator $\hat{Y}_{0M}$. Can be manipulated through model choice: becomes smaller as fewer variables are included in the model.
- Squared Bias: The last term is a bias term of $\hat{Y}_{0M}$ from the true expectation μ_0. Can be manipulated through model choice and becomes smaller as more variables are included in the model.

Bias-variance trade-off: The more complex the model, the smaller the squared bias and the greater the variance. On the contrary, simpler models show a greater squared bias and in return for a smaller variance.

Overfitting and Limit on Predictors in Classical Regression Analysis

Overfitting: when model is too complex (too many free parameters to estimate for the amount of information in the data), the goodness-of-fit of the model (e.g., R^2) will be exaggerated and future observed values will not agree with predicted values on independent samples.

To prevent overfitting, need to consider reliability or calibration of a model: the ability of the model to predict future observations as well as the observations collected.

Rule of thumb for classical regression models: Empirical studies show that a fitted regression model is likely to be reliable when the number of predictors (or candidate predictors if using variable selection) p is less than $m/10$ or $m/20$, where m is the "limiting sample size".

Overfitting and Limit on Predictors in Classical Regression Analysis

Limiting sample sizes m:

- continuous outcome: n (total sample size)
- Binary: $\min(n_1, n_2)$ (consider the power of a two-sample binomial test for effective sample size).
- Ordinal (k categories): $n - \frac{1}{n^2} \sum_{i=1}^{k} n_i^3$
- Time to event data: number of events

Note: Narrowly distributed predictor variables (e.g., if all subjects' ages are between 30 and 45 or only 5% of subjects are female) will require even higher sample sizes.

The number of candidate variables must include all variables screened for association with the response (e.g., nonlinear terms and interactions).

Model Selection

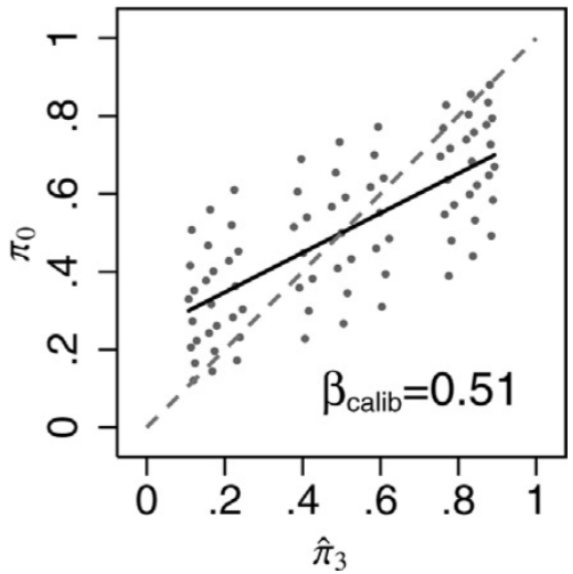
On the basis of scientific knowledge (e.g., literature), obtain a pre-selection of potential models.

All subset selection: assess all potential models by one of the various model choice criteria. The summary of the results should not be restricted to the "best" model. There could be a number of competitive models having approximately equal model fit, differing only in small aspects from each other. These differences cause some uncertainty regarding the conclusions.

In case that the number of covariates is smaller than about 40, we can determine the best model (in the sense of a model choice criterion) with the "leaps and bounds" algorithm. The algorithm returns the optimal model thereby avoiding the computation of all models.

Shrinkage

Shrinkage: the slope of a calibration plot (plot of observed responses against predicted responses) is less than one. When plot the calibration plot on the training data, slope is usually one. However, on independent samples, the slope (shrinkage factor) is less than one (low predictions will be too low and high predictions too high).



For OLS, an estimate of shrinkage factor is

$$\hat{\gamma} = \frac{n-p-1}{n-1}\frac{R^2_{adj}}{R^2}, \quad R^2_{adj} = 1-(1-R^2)\frac{n-1}{n-p-1}$$

Shrinkage

Shrinkage estimates to prevent overfitting: appropriately discounting $\hat{\boldsymbol{\beta}}$, make the model underfitted on the data at hand (i.e., apparent γ on training data < 1) so that on new data extremely low or high predictions are correct.

Ridge regression is one technique for placing restrictions on the parameter estimates that results in shrinkage. A ridge tuning parameter will be be chosen to control the amount of shrinkage.

Important how to choose tuning parameter: AIC, BIC, cross validation.

Penalized maximum likelihood estimation is a general shrinkage procedure. Penalty functions include LASSO, ridge, elastic net, SCAD and so on. Most effective for $p > n$ problems. It has been an active area of research in statistics during the past two decades.

Model Selection

Forward Selection: Based on a starting model, forward selection includes one additional variable in every iteration of the algorithm. The variable which offers the greatest reduction of a preselected model choice criteria (C_p, AIC, CV, BIC) is chosen. The algorithm terminates if no further reduction is possible.

Backward Elimination: starts with the full model containing all potential covariates. Subsequently, in every iteration, the covariate which provides the greatest reduction of the model choice criteria (C_p, AIC, CV, BIC) is eliminated from the model. The algorithm terminates if no further reduction is possible.

Stepwise Selection: A combination of forward selection and backward elimination. In every iteration of the algorithm, the inclusion and the deletion of a variable are both possible.

Issues with Forward, Backward, Stepwise Selection

- Some stepwise regression uses significance testing as criterion to determine model. But the test statistics do not have claimed distribution (distribution is derived for pre-specified hypothesis).
- Yields standard errors of regression coefficient estimates that are biased low and confidence intervals for effects and predicted values that are falsely narrow, p-values too small. inflated type-1 error
- Yields R^2 values that are biased high.
- Regression coefficients that are biased high in absolute value and need shrinkage.
- Rather than solving problems caused by collinearity, variable selection is made arbitrary by collinearity.
- Dummy variables of multi-categorical variables are neither jointly included nor removed from the model.
- Do not deal with hierarchical terms (e.g., polynomial terms or interaction terms).

Practical Guidelines and Alternatives

The problems of p-value-based variable selection are exacerbated when the results of the final model are interpreted as if it were prespecified (inflated type I error rate).

If stepwise selection must be used, use a global test: simultaneously testing all candidate predictors and having degrees of freedom equal to the number of candidate variables. If this global test is not significant, selection of individually significant predictors is usually not warranted.

Use alternatives: dimension reduction (e.g., PCA) or modern variable selection methods based on regularization (e.g., LASSO).

Model Validation

Model validation: ascertain whether predicted values from the model are likely to accurately predict responses on future subjects or subjects not used to develop our model.

Internal validation based on cross-validation: Randomly split one data into a training sample and a testing sample. Fit data on the training sample and evaluate performance on the testing sample.

External validation: Develop the model on one study sample. Validate the model on an independent sample.

Three major causes of failure: overfitting, changes in measurement methods/changes in definition of categorical variables, and major changes in subject inclusion criteria.

Model Validation

In OLS, model validation can be based on R^2: compare R^2 in the training sample to that achieved in the test sample. A drop in R^2 indicates overfitting, and the absolute R^2 in the test sample is an unbiased estimate of predictive discrimination.

In extremely overfitted models, R^2 in the test set can be negative, since it is computed on fixed intercept and regression coefficients using the formula 1 - SSE/SST, where SSE is the error sum of squares, SST is the total sum of squares, and SSE can be greater than SST (when predictions are worse than the constant predictor $\bar{Y}$).

k-fold cross validation (CV) is often used. Repeat B times.

Estimate overfitting by the randomization method: "How well can the response be predicted when we use our best procedure on random responses when the predictive accuracy should be near zero?" The better the fit on random Y, the worse the overfitting.

Summary on Developing Prediction Models

- Specify candidate predictors (e.g., based on literature)
- Examine missing data and take appropriate action (e.g., impute X, drop predictor with substantial missing)
- For each predictor, consider transformation or whether nonlinear relationship is needed. Consider interaction terms.
- If the number of terms fitted or tested in the modeling process (counting nonlinear and interaction terms) is too large in comparison with the sample size, consider dimension reduction.
- Consider structured variables jointly (e.g., test all dummy variables for a categorical predictor simultaneously).
- Check model assumptions (e.g., additivity, interactions, influential points, distribution assumptions)
- Variable selection if parsimony is more important than accuracy

Summary on Developing Prediction Models

- Interpret the model
- Validate the final model for calibration and discrimination ability (e.g., cross validation/bootstrap).
- When missing values were imputed, adjust final variance-covariance matrix for imputation (e.g., using bootstrap or multiple imputation).
- Report results

Modern Techniques for High-dimensional Data Analysis

High-dimensional Data Analysis

Large scale data: large sample size, large number of predictors

- Data collected on hundreds of thousands or millions of subjects with a diverse array of variables
- Intensive time series of biologic signals collected every few milliseconds
- Extremely large data arrays where almost all the variables are of one type

Examples:

- neuroimaging
- SNPs for genome-wide association studies
- RNA sequencing
- wearable devices
- electronic health records

High-dimensional Data Analysis

Findings from high-dimensional biological data are often inconsistent

- Biology is complex
- Use of non-reproducible research methodology
- Unknown properties of statistical methods; statistical theoretical results developed under unrealistic (or unverifiable) conditions
- Multiple comparison problems and double dipping
- Inadequate sample size for the complexity of the analytic task

Need more research to understand statistical inference of modern techniques

High-dimensional Data Analytic Approaches

Regularized (penalized) regression:

- lasso: a penalty on the absolute value of regression coefficient that highly favors zero as an estimate. This results in a large number of estimated coefficients being exactly zero, i.e., results in feature selection. However, selected variables may be unstable (use stability selection)
- ridge regression (penalty function that is quadratic in the regression coefficients): does not result in a parsimonious model but can have a high predictive value
- elastic net: a combination of lasso and quadratic penalty that has some parsimony but has better predictive ability than the lasso. The difficulty is simultaneously choosing two penalty parameters (one for absolute value of βs, one for their sum of squares).

High-dimensional Data Analytic Approaches

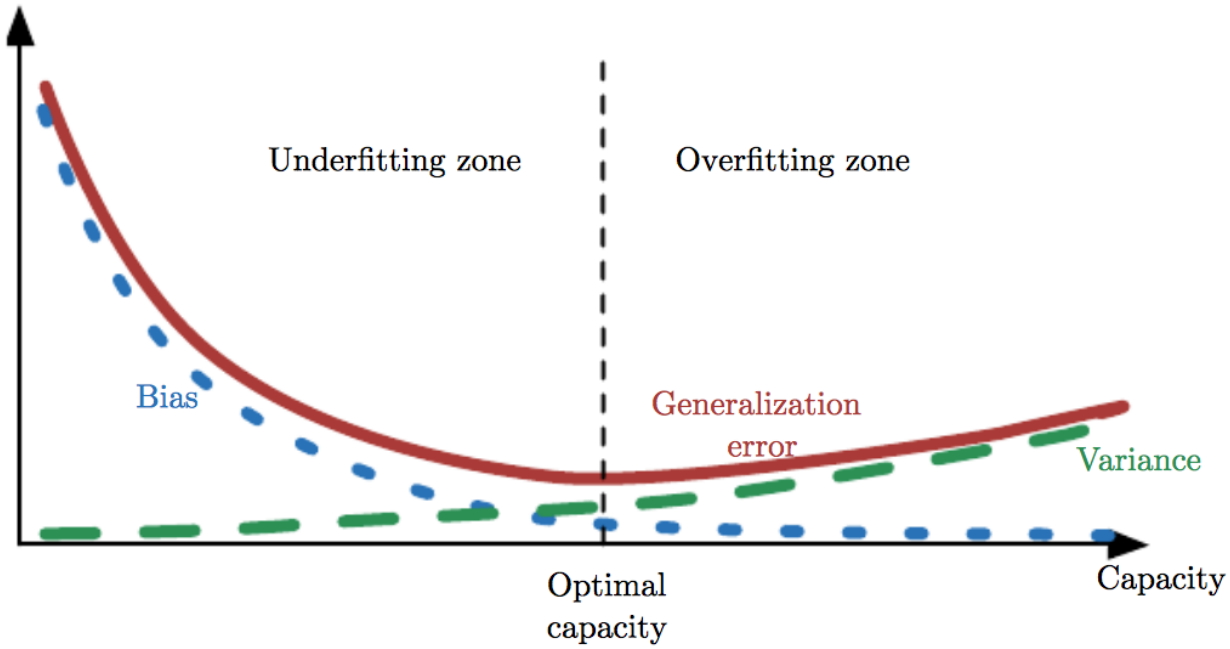
Random forest: fit a regression tree using recursive partitioning (CART) on multiple random samples of candidate features. Multiple trees are combined. The result is no longer a tree; it is uninterpretable. Often competitive in predictive ability.

Dimension reduction followed by traditional regression modeling (PCA regression).

Incorporate biological information into high-dimensional data analysis through designing an appropriate penalty function (pathway information of genes, structured variable selection, network information).

Bias Variance Trade-Off

$$\text{MSE}=E[(\hat{\beta}_m - \beta)^2]=[\text{Bias}(\hat{\beta}_m)]^2+\text{Var}(\hat{\beta}_m)$$



As model complexity/capacity increases (x-axis), bias (dotted) tends to decrease and variance (dashed) tends to increase, yielding a U-shaped curve for generalization error.

Cross validation to find the 'sweet spot'.

Emerging Topics: Modeling Brain or Symptom Networks

Modeling Networks

Why modeling brain as a complex system through the lens of networks/connectomics?

- Brain regions do not act in isolation
- Interconnections are characteristics of brain activities
- Network disruption is often associated with brain disorders (mental disorders, neurological disorders, aging)

Network Analysis

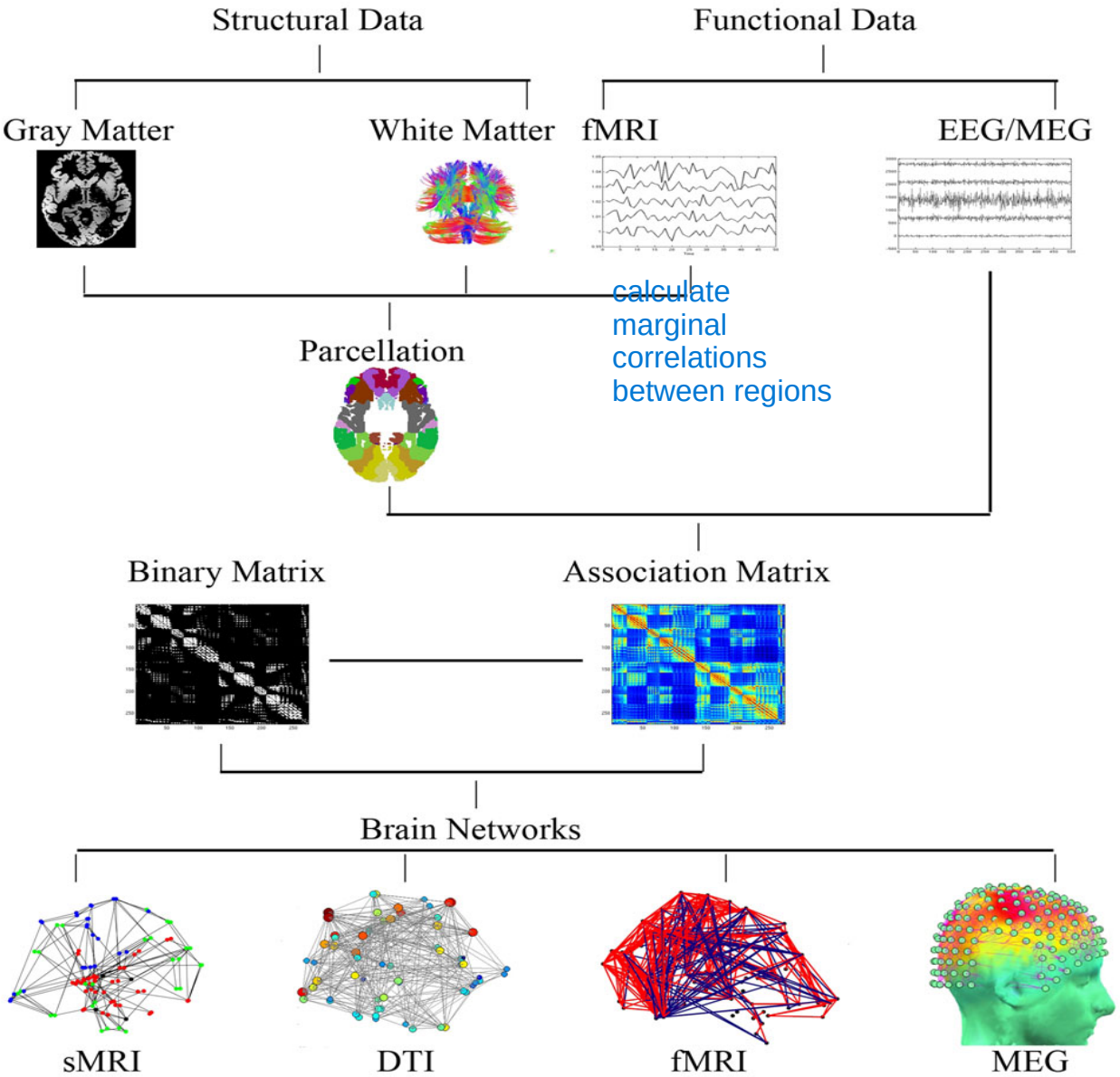
Networks are represented by graphs. Network analysis treats nodes in a network (i.e., biomarkers) as a system and investigate the interrelationships (interaction, joint distribution) between nodes.

Need to define nodes and edges in a graph:

- Nodes: regions in the brain according to anatomical parcellations or functional parcellations (brain network); gene expressions (gene network); subjects (social network)
- Edges: association between nodes (correlation, partial correlation); social interactions (friendship in a social media network)

Analytic tools: biological experiments; statistical association analysis (e.g., correlation between time series, Gaussian graphical model to estimate precision matrix)

Brain Networks



Bassett & Bullmore (2010). *Curr Op Neurol*

Case Study: Dynamic Gaussian Graphical Models