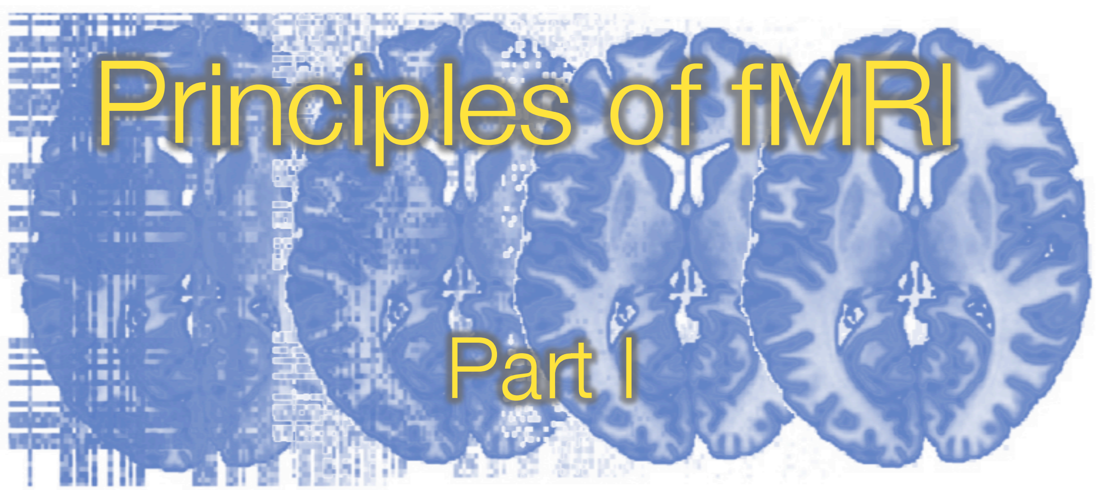
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

University of Colorado **Boulder**

CANlab

SMART JHSPH Biostatistics



Martin Lindquist
Department of Biostatistics
Johns Hopkins
Bloomberg School of Public Health

Tor Wager
Department of Psychology and
Neuroscience and the
Institute for Cognitive Science
University of Colorado, Boulder

@fMRIstats

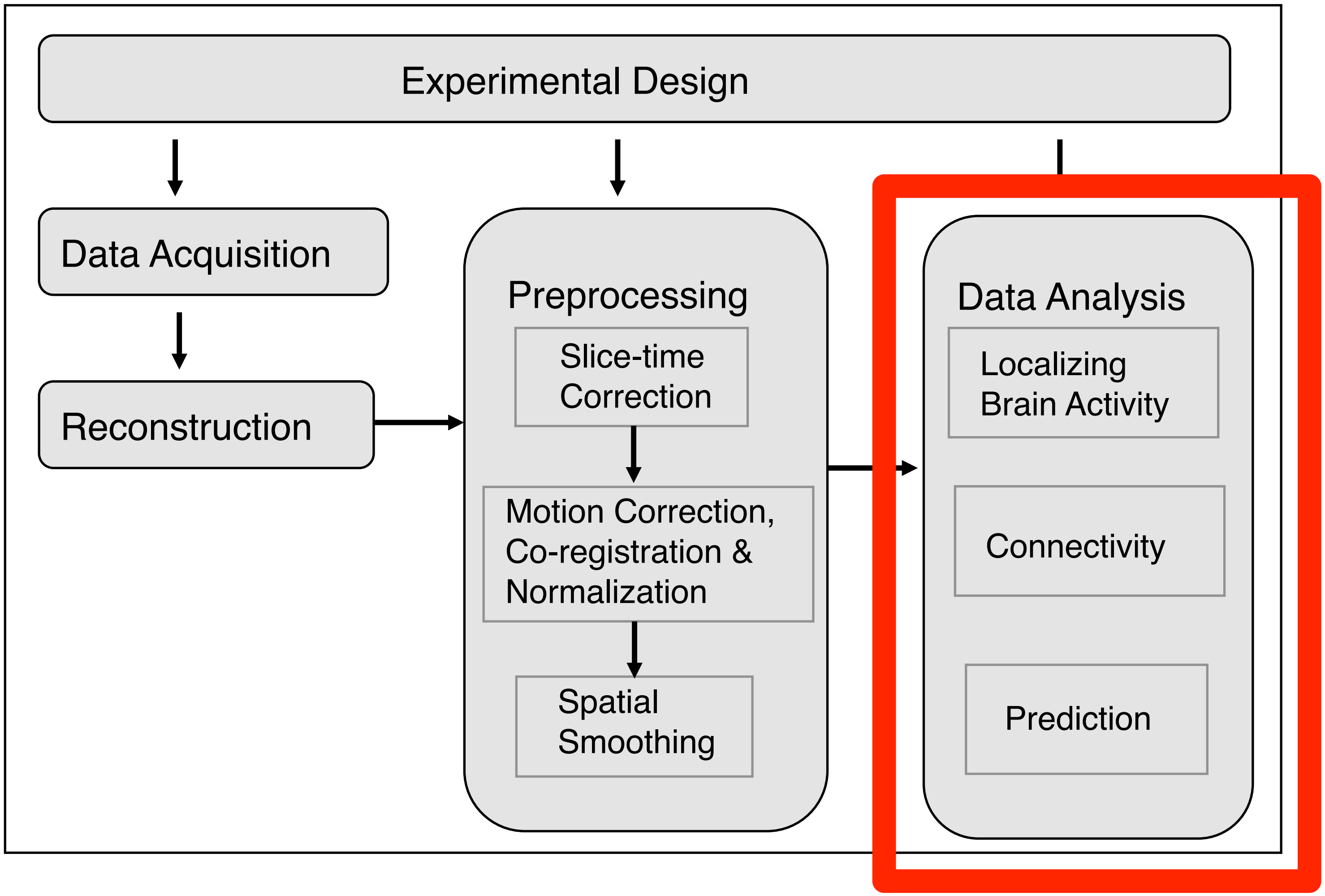
The General Linear Model: Introduction

Statistical Analysis

- There are multiple goals in the statistical analysis of fMRI data.

- They include:
  - localizing brain areas activated by the task;
  - determining networks corresponding to brain function; and
  - making predictions about psychological or disease states.

Data Processing Pipeline

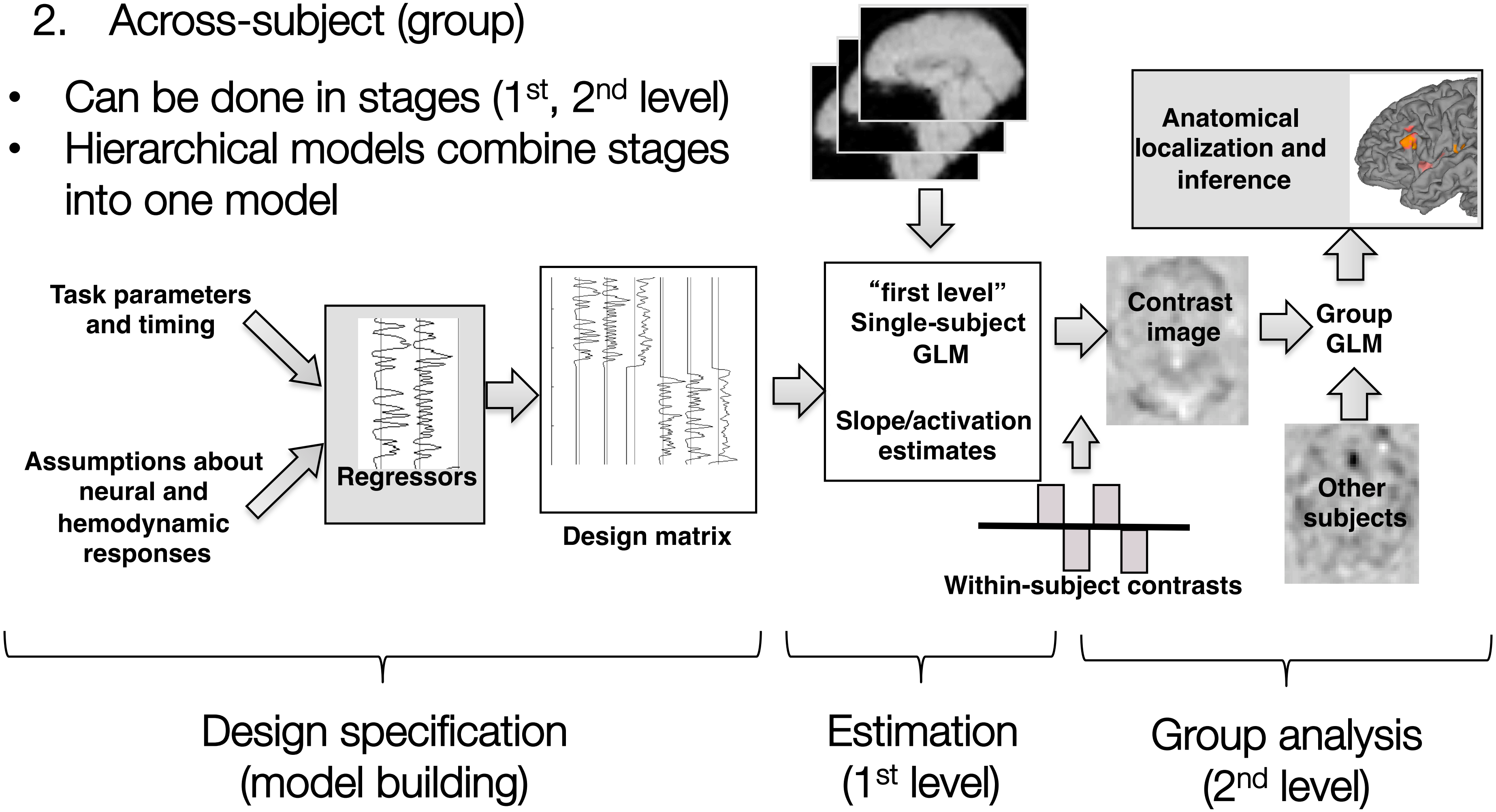


Overview of the GLM analysis process

Typically a two-level hierarchical analysis

1. Within-subject (individual)
2. Across-subject (group)

- Can be done in stages (1st, 2nd level)
- Hierarchical models combine stages into one model



INTRODUCING THE GLM FAMILY

@fMRIstats

Lindquist & Wager

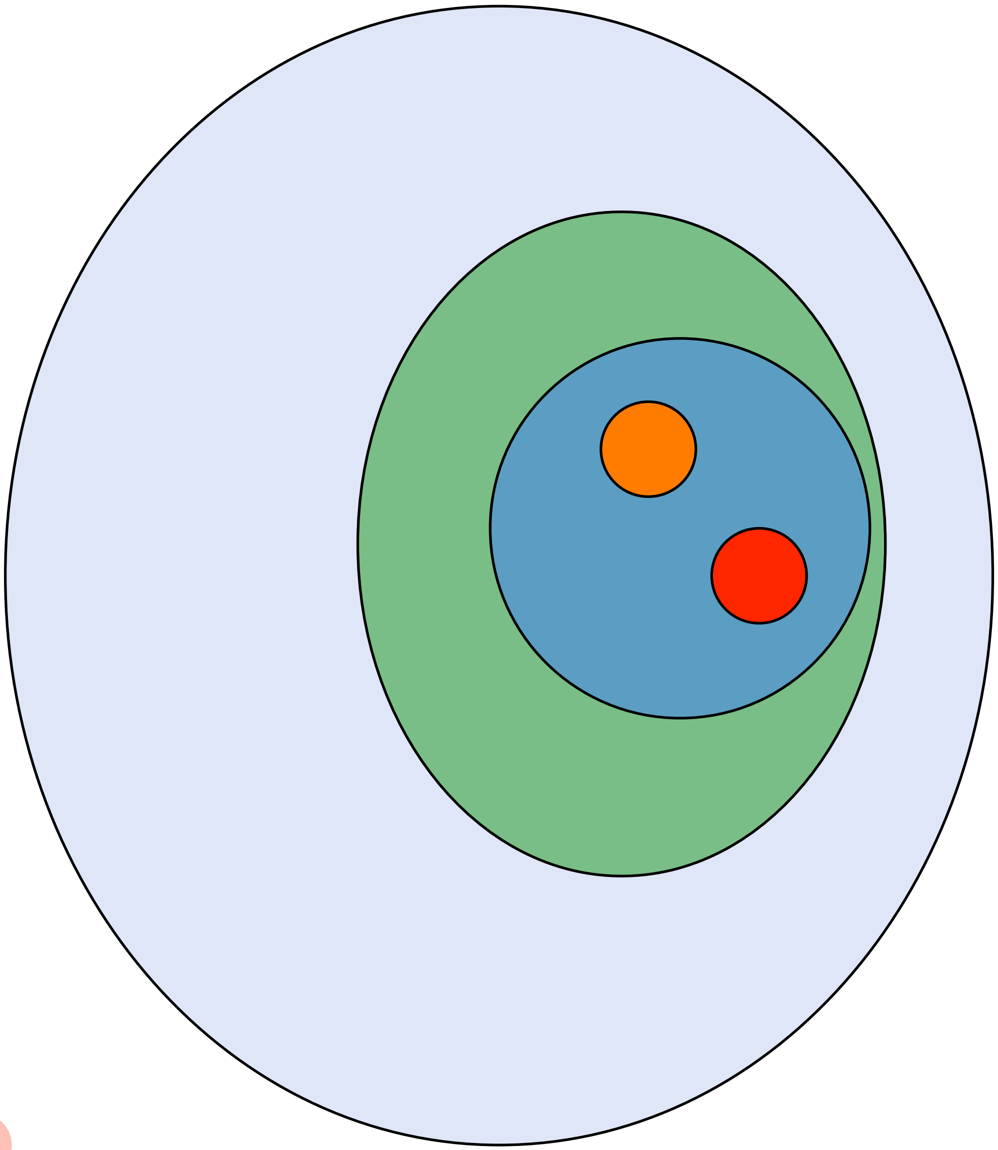
General Linear Model

- The general linear model (GLM) approach treats the data as a linear combination of model functions (predictors) plus noise (error).

- The model functions are assumed to have known shapes (i.e., straight line, or known curve), but their amplitudes (i.e., slopes) are unknown and need to be estimated.

- The GLM framework encompasses many of the commonly used techniques in fMRI data analysis (and data analysis more generally).

The GLM Family



Simple regression

ANOVA

Multiple regression

General linear model

- Mixed effects/hierarchical
- Timeseries models (e.g., autoregressive)
- Robust, penalized regression (LASSO, Ridge)

Generalized linear models

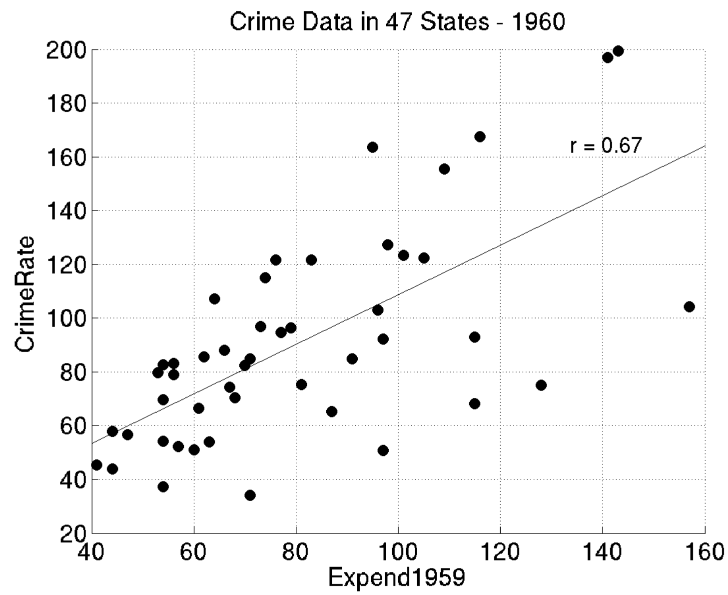
- Non-normal errors
- Binary/categorical outcomes (logistic regression)

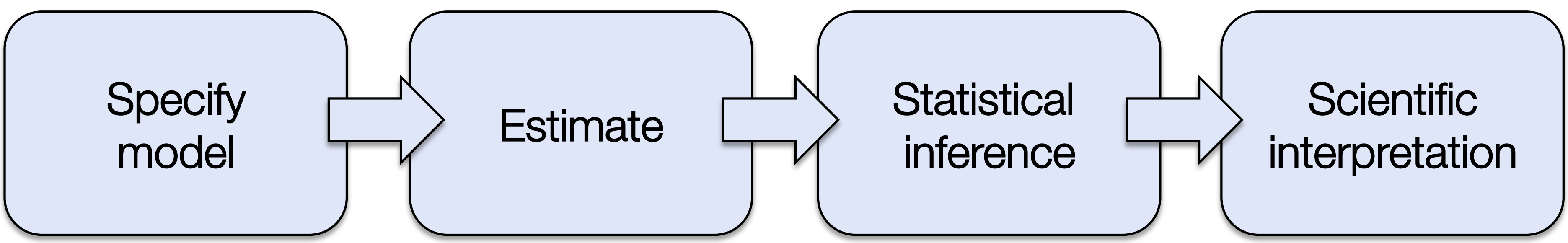
Closed form solution

Iterative solutions

Simple regression:
A simple linear model with one predictor

Principles of fMRI
Part I





Simplification: "linear relationship"

Find slope, intercept

Test slope: P-value

Meaning of relationship?

The GLM Family

Dependent Variable	Predictors	Analysis
One continuous	Continuous One predictor	Regression
	Continuous Two+ preds	Multiple Regression
	Categorical 1 pred., 2 levels	2-sample t-test
	Categorical 1 p., 3+ levels	One-way ANOVA
	Categorical 2+ predictors	Factorial ANOVA

General Linear Model

The GLM Family

Dependent Variable	Predictors	Analysis
One continuous	Continuous 1+	Multiple Regression
	Categorical 1+	ANOVA
	Categorical Continuous	ANCOVA
2 repeated measures	0 between 1 within	Paired t-test
k repeated measures	0 between k – 1 categorical within	One-way RM-ANOVA, MANOVA
4+ repeated measures	0 between *2+ categorical* within	Factorial RM-ANOVA
4+ repeated measures	1+ between *2+ categorical* within	"GLM"

General Linear Model

Generalized least squares (iterative): correlated errors (e.g., timeseries)

Correlated errors across repeated measures

The multiple regression model
basic model for the GLM

- Structural model for regression

Variables: DV, Pred1, Pred2, Pred*k*

$$\overset{\text{DV}}{y_i} = \underset{\text{intercept}}{\beta_0} + \underset{\text{Slope 1}}{\beta_1} \overset{\text{Pred1}}{x_{i1}} + \underset{\text{Slope 2}}{\beta_2} \overset{\text{Pred2}}{x_{i2}} + \cdots + \underset{\text{Slope k}}{\beta_k} \overset{\text{Pred}k}{x_{ik}} + \underset{\text{Error}}{\varepsilon_i}$$

Parameters: intercept, Slope 1, Slope 2, Slope k, Error

- solve for beta vector
- minimize sum of squared residuals

Matrix notation

$$y = X\beta + \varepsilon$$

Matrix Notation

- Alternatively, we can write

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

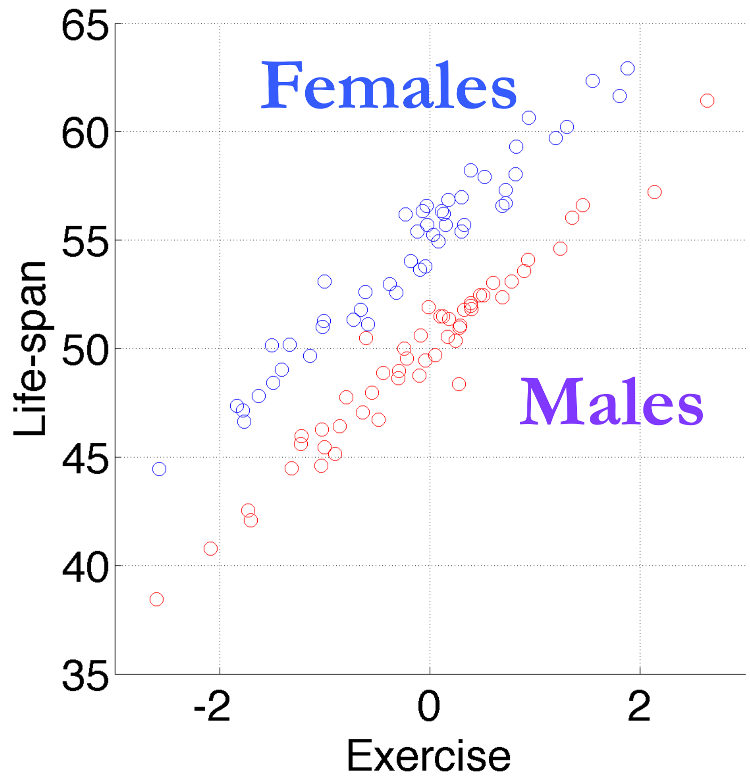
as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{np} & \cdots & X_{np} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Observed Data **Design matrix** **Model parameters** **Residuals**

A non-fMRI example

- Does exercise predict life-span?
- Made-up (not real data)
- Control for other variables that might be important, i.e., gender (M/F)

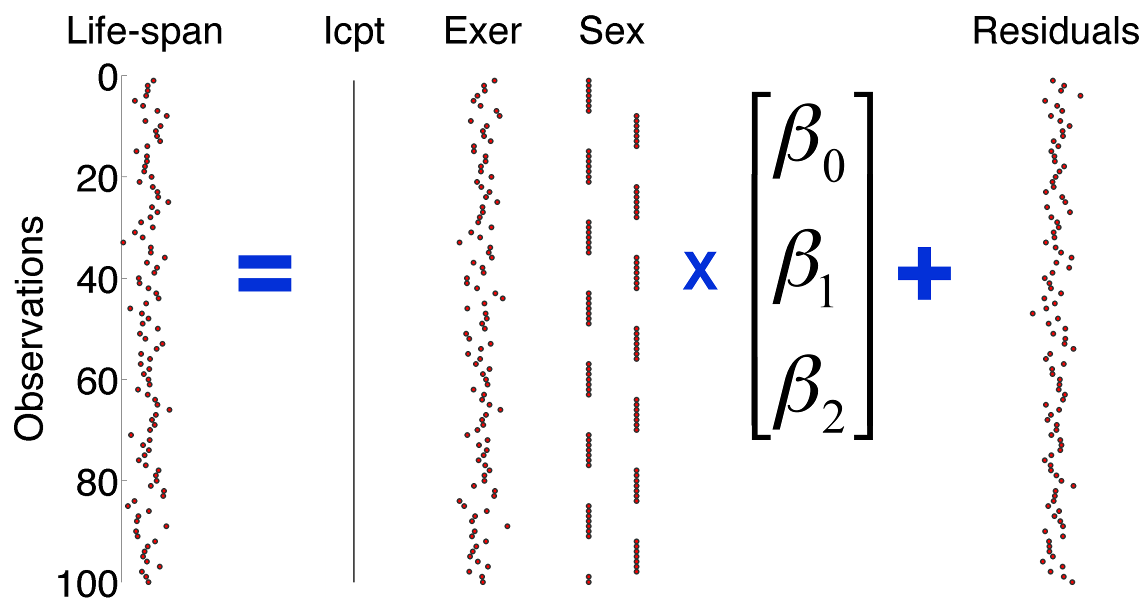


A non-fMRI example

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

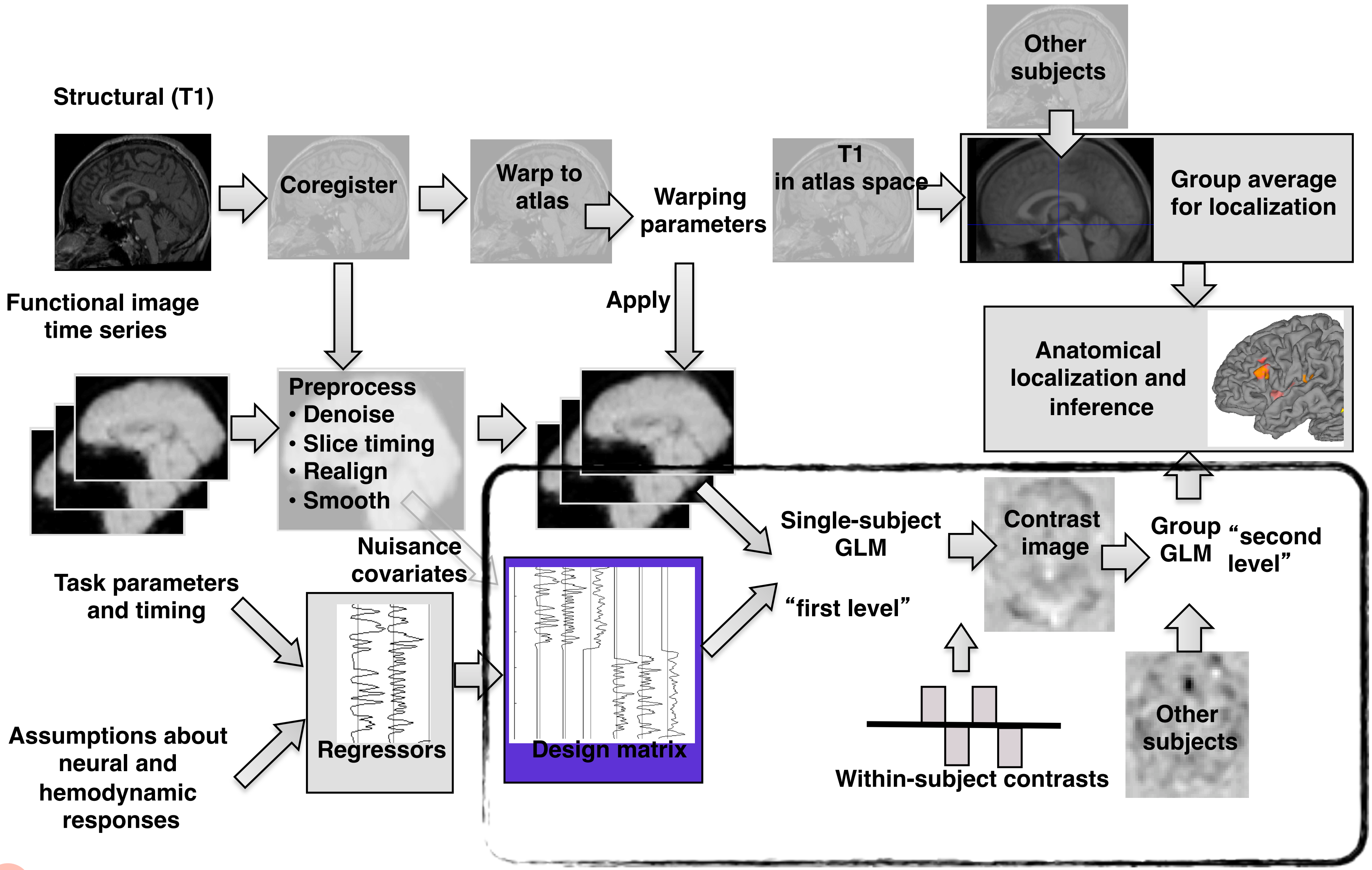
$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{np} & \cdots & X_{np} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Outcome Data **Design matrix** **Model parameters** **Residuals**



End of Module

A standard processing pipeline



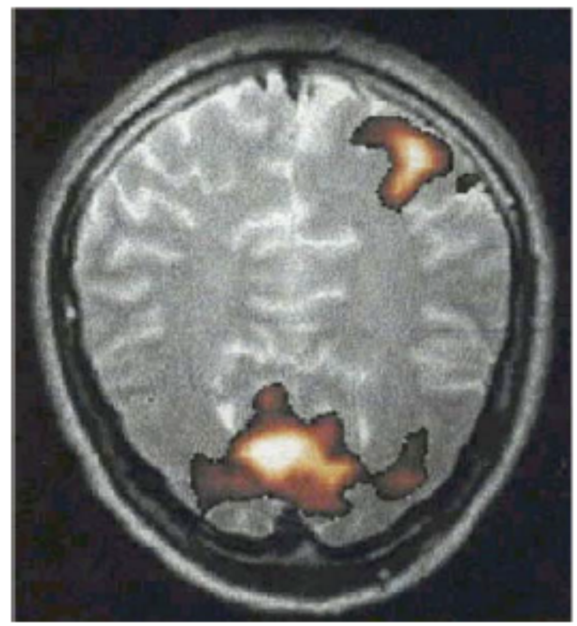
General Linear Model

- General model
  - Categorical or continuous predictor
- Linear model
  - Parameters that are
    - not multiplied by other parameters, e.g., not $Y = abX$
    - only first-power, e.g., NOT $Y = b^2X$
    - not exponents, e.g., NOT $Y = X^b$
  - Variables do not need to satisfy the above criteria
    - transformation workaround, e.g., $Y = bX^2$ can be rewritten as $Y = bZ$
  - Not necessary straight-line relationship

Human Brain Mapping

- The most common use of fMRI to date has been to localize areas of the brain that activate in response to a certain task.

- These types of human brain mapping studies are necessary for the development of biomarkers and increasing our understanding of brain function.



The regression model

- Vector notation: y, x, and e are vectors of N values corresponding to N observations

$$y = b + mx + e$$

Outcome (DV)

Intercept (constant)

slope

Predictor value

Error (residual)

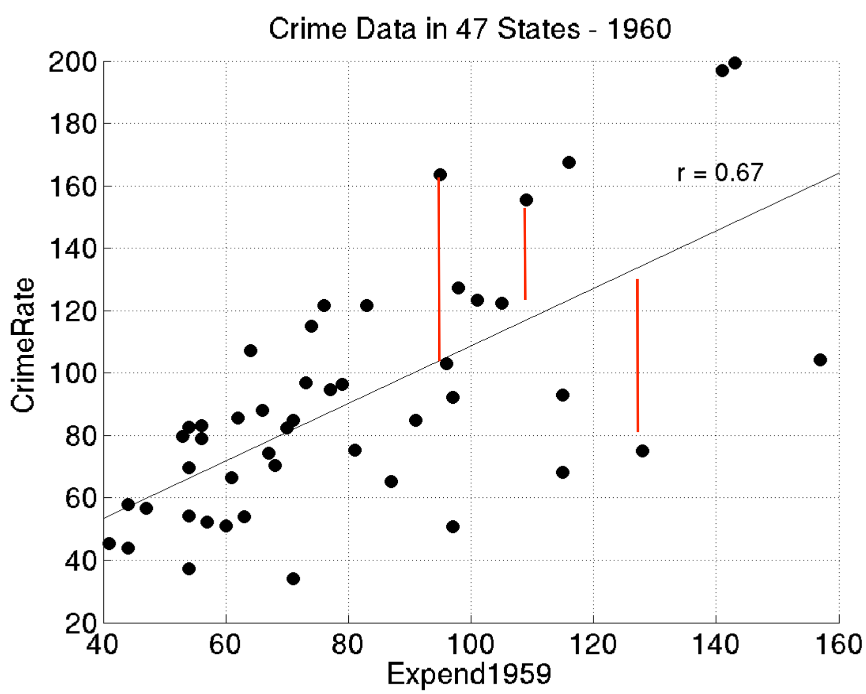
For point i: $$y_i = b + mx_i + e_i$$

Fitting and residual variance

- Minimize squares of vertical distance to lines.

  Minimize: $\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$

  – The line is being pulled by vertical rubber bands attached to each point. Vector of red lines is *e*



Matrix Notation

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{np} & \cdots & X_{np} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Observed Data **Design matrix** **Model parameters** **Residuals**

Is the same as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \beta_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{np} \end{bmatrix} + \beta_p \begin{bmatrix} X_{1p} \\ X_{2p} \\ \vdots \\ X_{np} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Design matrix

- In fMRI the design matrix specifies how the factors of the model change over time.

- The design matrix is an $n \times p$ matrix where n is the number of observations over time and p is the number of model parameters

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{np} & \cdots & X_{np} \end{bmatrix}$$