

Introduction to Linear Models

STA721 Linear Models Duke University

Merlise Clyde

August 25, 2015

Coordinates

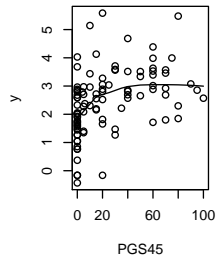
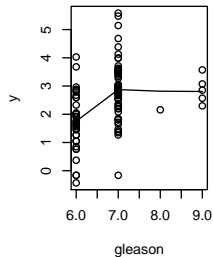
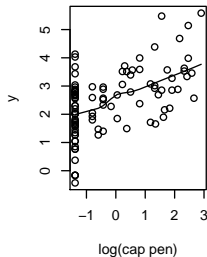
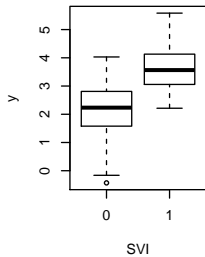
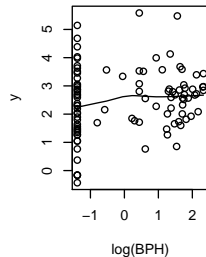
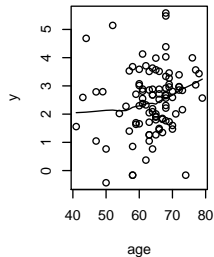
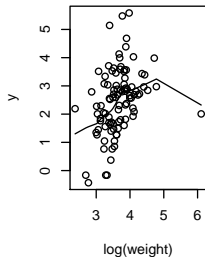
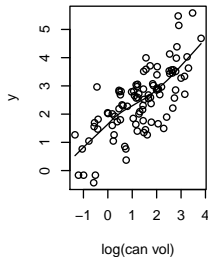
- ▶ Instructor: Merlise Clyde
214 Old Chemistry
Office Hours MWF 1:00-2:0 or right after class (or by appointment)
- ▶ Teaching Assistants: Nicole Dalzell & Shin Shirota
- ▶ Course: Theory and Application of linear models from both a frequentist (classical) and Bayesian perspective
- ▶ Prerequisites: linear algebra and a mathematical statistics course covering likelihoods and distribution theory (normal, t, F, chi-square, gamma distributions)
- ▶ Introduce R programming as needed
- ▶ Introduce Bayesian methods, but assume that you are co-registered in 601 or have taken it previously
- ▶ more info on Course Website
<http://stat.duke.edu/courses/Fall15/sta721>

Introduction

Build “regression” models that relate a response variable to a collection of covariates

- ▶ Goals of Analysis?
 - ▶ Predictive models
 - ▶ Causal interpretation
 - ▶ Testing of hypotheses
 - ▶ confirmatory or validation analyses
- ▶ Observational versus Experimental data? (Confounding)
- ▶ Sampling Schemes Generalizability
- ▶ Statistical Theory

Prostate Example



Simple Linear Regression

Simple Linear Regression:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ for } i = 1, \dots, n$$

Rewrite in vectors:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \beta_1 + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Multiple Regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots \beta_p x_{pi} + \epsilon_i$$

Design matrix

$$\mathbf{X} = \begin{matrix} & 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \dots & x_{pn} \end{matrix}$$

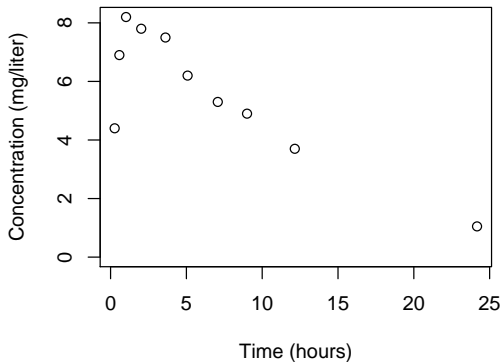
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

what should go into \mathbf{X} and do we need all columns of \mathbf{X} for inference about \mathbf{Y} ?

Nonlinear Models

Mean function may be an intrinsically nonlinear function of t

$$E[Y_i] = f(t_i, \theta)$$



Quadratic Linear Regression

Taylor's Theorem:

$$f(t_i, \boldsymbol{\theta}) = f(t_0, \boldsymbol{\theta}) + (t_i - t_0)f'(t_0, \boldsymbol{\theta}) + (t_i - t_0)^2 \frac{f''(t_0, \boldsymbol{\theta})}{2} + R(t_i, \boldsymbol{\theta})$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \text{ for } i = 1, \dots, n$$

Rewrite in vectors:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Quadratic in x , but linear in β 's, but remainder term is in errors ϵ

Polynomial Linear Regression

Polynomial Regression:

$$y_i = \sum_{j=0}^q \beta_j x_i^j + \epsilon_i \text{ for } i = 1, \dots, n$$

Rewrite in vector notation:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^q \\ \vdots & \vdots & & & \\ 1 & x_n & x_n^2 & \dots & x_n^q \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Y =

Xβ + ε

How large should q be?

Use Nonlinear Regression or other Nonparametric models

Kernel Regression

Kernel Regression:

$$y_i = \beta_0 + \sum_{j=1}^J \beta_j e^{-\lambda(x_i - k_j)^d} + \epsilon_i \text{ for } i = 1, \dots, n$$

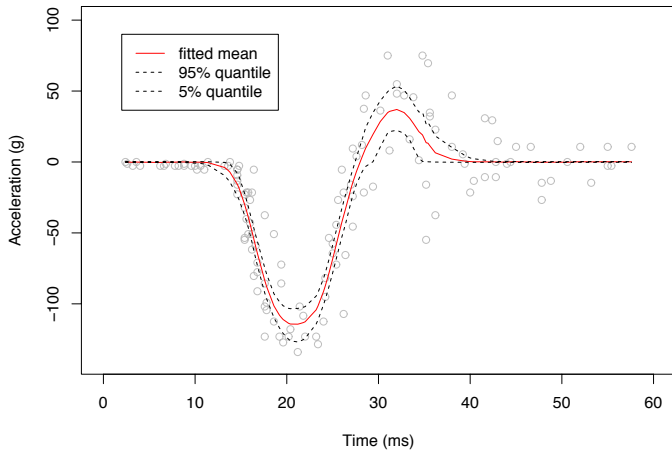
where k_j are kernel locations and λ is a smoothing parameter

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & e^{-\lambda(x_1 - k_1)^d} & \dots & e^{-\lambda(x_1 - k_J)^d} \\ \vdots & \vdots & & \vdots \\ 1 & e^{-\lambda(x_n - k_1)^d} & \dots & e^{-\lambda(x_n - k_J)^d} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_J \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

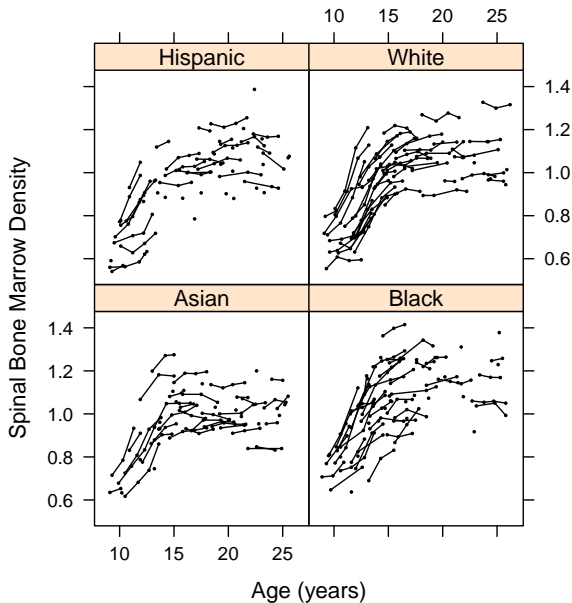
Linear in $\boldsymbol{\beta}$ given λ

Learn λ and J

Kernel Regression Example



Hierarchical Models - Spinal Bone Density



Generic Linear Model

Generic Model in Matrix Notation is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- ▶ \mathbf{Y} ($n \times 1$) vector of response (observe)
- ▶ \mathbf{X} ($n \times p$) design matrix (observe)
- ▶ $\boldsymbol{\beta}$ ($p \times 1$) vector of coefficients (unknown)
- ▶ $\boldsymbol{\epsilon}$ ($n \times 1$) vector of “errors” (unobservable)

Goals:

- ▶ What goes into \mathbf{X} ? (model building and model selection)
- ▶ What if several models are equally good? (model averaging)
- ▶ What about the future? (Prediction)
- ▶ uncertainty quantification - assumptions about $\boldsymbol{\epsilon}$

All models are wrong, but some may be useful (George Box)

Ordinary Least Squares

Goal: Find the best fitting “line” or “hyper-plane” that minimizes

$$\sum_i (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

- ▶ Optimization problem
- ▶ May over-fit \Rightarrow add other criteria that provide a penalty
“Penalized Least Squares”
- ▶ Robustness to extreme points \Rightarrow replace quadratic loss with other functions
- ▶ no notion of uncertainty of estimates
- ▶ no structure of problem (repeated measures on individual, randomization restrictions, etc)

Need Distribution Assumptions of Y (or ϵ) for testing and uncertainty measures \Rightarrow Likelihood and Bayesian inference

Philosophy

- ▶ for many problems frequentist and Bayesian methods will give similar answers (more a matter of taste in interpretation)
- ▶ For small problems, Bayesian methods allow us to incorporate prior information which provides better calibrated answers
- ▶ for problems with complex designs and/or missing data Bayesian methods are often easier to implement (do not need to rely on asymptotics)
- ▶ For problems involving hypothesis testing or model selection frequentist and Bayesian methods can be strikingly different.
- ▶ Frequentist methods often faster (particularly with “big data”) so great for exploratory analysis and for building a “data-sense”
- ▶ Bayesian methods sit on top of Frequentist Likelihood

Important to understand advantages and problems of each perspective!