# Nonlinear Regression

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

# Chapter 13 Parametric nonlinear regression

Throughout most of STAT 704 and 705, we concentrated on linear models where $E(Y_i) = \mathbf{x}_i'\boldsymbol{\beta}$. Notable exceptions arose when we considered non-normal data. For logistic regression we had $E(Y_i) = e^{\mathbf{x}_i'\boldsymbol{\beta}}/[1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}]$; Poisson regression gave us $E(Y_i) = t_i e^{\mathbf{x}_i'\boldsymbol{\beta}}$.

Sometimes scientists have a parametric non-linear mean function in mind for normal data. Theoretical considerations may lead to such a model, or else empirical evidence collected over time. Examples: dose-response models, growth curves, heating in swine due to MRI.

# Parametric nonlinear regression

A parametric nonlinear model (13.1–13.5) has a prespecified parametric form indexed by parameters $\gamma$

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\gamma}) + \epsilon_i.$$

For example the exponential growth/decay model is $Y_i = \gamma_0 e^{\gamma_1 x_i} + \epsilon_i$. Data reduction takes place through the estimation of $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ and $\sigma$.

Other examples are the logistic growth curve $Y_i = \gamma_0[1 + \gamma_1 \exp(\gamma_2 x_i)]^{-1} + \epsilon_i$ and the von Bertlanffy growth curve $Y_i = L_\infty \left[1 - \exp\left(-K(x_i - x_0)\right)\right] + \epsilon_i$.

Note that model diagnostics are similar to the linear case, for example $r_i = Y_i - f(\mathbf{x}_i, \hat{\gamma})$ can be used to assess model adequacy.

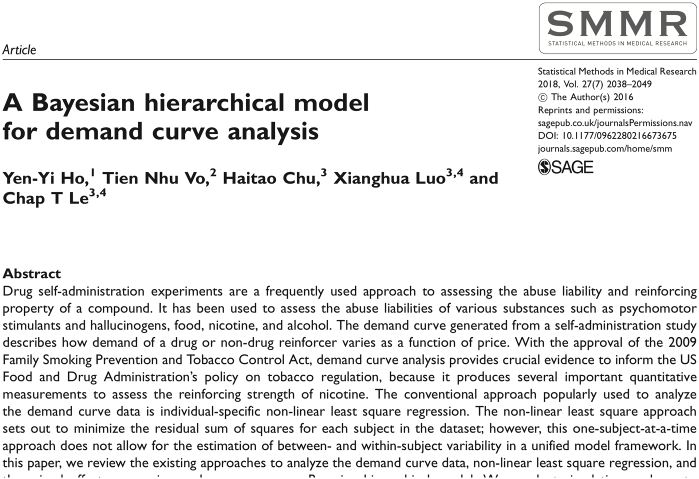
# Fitting parametric nonlinear models

Fitting of such models is carried out via maximum likelihood using Newton-Raphson. Several functions in SAS can carry this out; PROC NLMIXED is the most versatile, while PROC NLIN is the old-school workhorse. Good starting values can make or break the program (as we'll see); you need to think about what the parameters represent in the model.

There is a bit on fitting at the end of the logistic regression notes. In your book see pp. 517–521. This theory is covered in more detail in STAT 823 (large sample theory) and STAT 740 (advanced statistical computing).

PROC NLMIXED provides the MLE's as well as standard errors. Also, functions of parameters can be estimated as well.

# Example: Demand Curve Analysis



*Article*

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2018, Vol. 27(7) 2038–2049
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280216673675
journals.sagepub.com/home/smm
SAGE

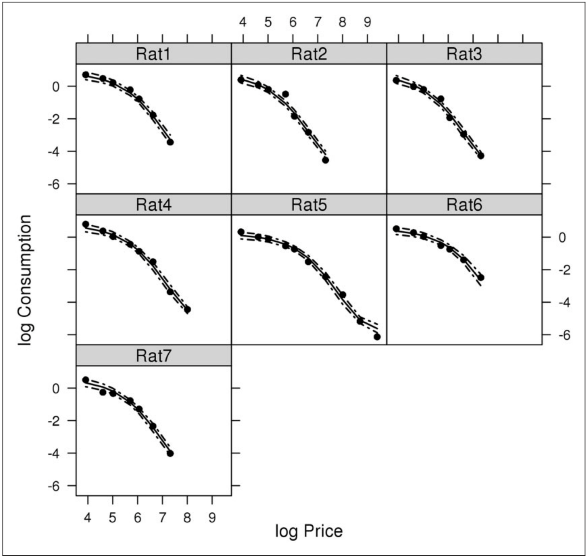
## A Bayesian hierarchical model for demand curve analysis

**Yen-Yi Ho,[1] Tien Nhu Vo,[2] Haitao Chu,[3] Xianghua Luo[3,4] and Chap T Le[3,4]**

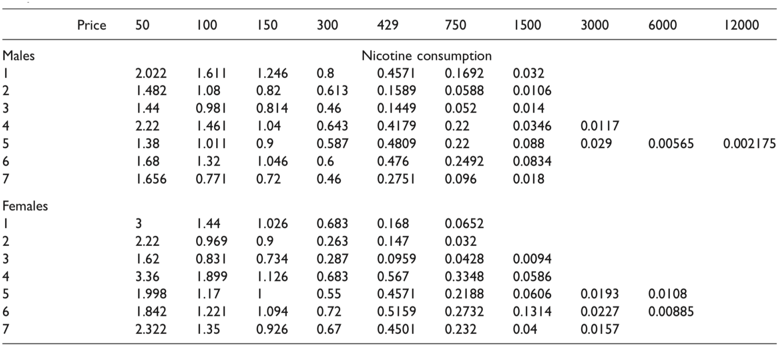
### Abstract

Drug self-administration experiments are a frequently used approach to assessing the abuse liability and reinforcing property of a compound. It has been used to assess the abuse liabilities of various substances such as psychomotor stimulants and hallucinogens, food, nicotine, and alcohol. The demand curve generated from a self-administration study describes how demand of a drug or non-drug reinforcer varies as a function of price. With the approval of the 2009 Family Smoking Prevention and Tobacco Control Act, demand curve analysis provides crucial evidence to inform the US Food and Drug Administration's policy on tobacco regulation, because it produces several important quantitative measurements to assess the reinforcing strength of nicotine. The conventional approach popularly used to analyze the demand curve data is individual-specific non-linear least square regression. The non-linear least square approach sets out to minimize the residual sum of squares for each subject in the dataset; however, this one-subject-at-a-time approach does not allow for the estimation of between- and within-subject variability in a unified model framework. In this paper, we review the existing approaches to analyze the demand curve data, non-linear least square regression, and

# Example: Demand Curve Analysis

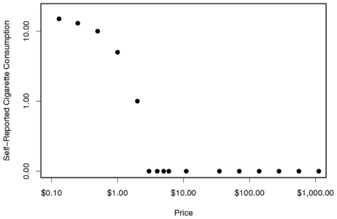


# Example: Demand Curve Analysis



| Price | 50 | 100 | 150 | 300 | 429 | 750 | 1500 | 3000 | 6000 | 12000 |
|---|---|---|---|---|---|---|---|---|---|---|
| Males | | | | | Nicotine consumption | | | | | |
| 1 | 2.022 | 1.611 | 1.246 | 0.8 | 0.4571 | 0.1692 | 0.032 | | | |
| 2 | 1.482 | 1.08 | 0.82 | 0.613 | 0.1589 | 0.0588 | 0.0106 | | | |
| 3 | 1.44 | 0.981 | 0.814 | 0.46 | 0.1449 | 0.052 | 0.014 | | | |
| 4 | 2.22 | 1.461 | 1.04 | 0.643 | 0.4179 | 0.22 | 0.0346 | 0.0117 | | |
| 5 | 1.38 | 1.011 | 0.9 | 0.587 | 0.4809 | 0.22 | 0.088 | 0.029 | 0.00565 | 0.002175 |
| 6 | 1.68 | 1.32 | 1.046 | 0.6 | 0.476 | 0.2492 | 0.0834 | | | |
| 7 | 1.656 | 0.771 | 0.72 | 0.46 | 0.2751 | 0.096 | 0.018 | | | |
| Females | | | | | | | | | | |
| 1 | 3 | 1.44 | 1.026 | 0.683 | 0.168 | 0.0652 | | | | |
| 2 | 2.22 | 0.969 | 0.9 | 0.263 | 0.147 | 0.032 | | | | |
| 3 | 1.62 | 0.831 | 0.734 | 0.287 | 0.0959 | 0.0428 | 0.0094 | | | |
| 4 | 3.36 | 1.899 | 1.126 | 0.683 | 0.567 | 0.3348 | 0.0586 | | | |
| 5 | 1.998 | 1.17 | 1 | 0.55 | 0.4571 | 0.2188 | 0.0606 | 0.0193 | 0.0108 | |
| 6 | 1.842 | 1.221 | 1.094 | 0.72 | 0.5159 | 0.2732 | 0.1314 | 0.0227 | 0.00885 | |
| 7 | 2.322 | 1.35 | 0.926 | 0.67 | 0.4501 | 0.232 | 0.04 | 0.0157 | | |

# The Hursh-Sylberberg Model



$$\log Q = \log Q_0 + k(e^{-\alpha P} - 1)$$

- $P$ is Price
- $Q$ is Demand/Consumption
- $Q_0$ is Level of demand when price approaches 0
- $k$ is related to the range of Q
- $\alpha$ is a measure of elasticity: the rate of decline in relative log consumption
- $P_{max}$: is the corresponding unit price.
- $O_{max}$: is the maximum expense (price times consumption) a person is willing to spend for a commodity

# Scientific Question

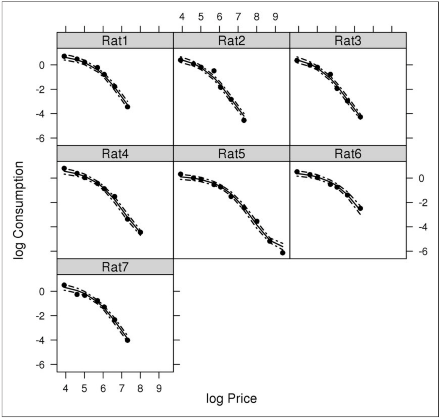
- Is the relative reinforcing efficacy (RRE) the same for female and male rats?
- Translate: Are $\alpha$, $Q_{max}$, $P_{max}$ the same for male and female rats?

# Various Approaches

- Individual-specific non-linear least squares regression model (NLIN)
- Mixed-effects model
- Bayesian Hierarchical Model

# The non-linear least squares regression model (NLIN)

- Estimating K based on ($\log_e Q_{max} - \log_e Q_{min}$) and set as a constant for all individuals
- Estimating $\alpha$ and $Q_0$ by minimizing the residual sum of squares, one-individual-at-a-time.



# NLIN: Pros and Cons

- Cons

# NLIN: Pros and Cons

- Cons
  - The parameters are estimated using individual data; not all data.

# NLIN: Pros and Cons

- Cons
  - The parameters are estimated using individual data; not all data.
  - The estimate for K is questionable.

# NLIN: Pros and Cons

- Cons
  - The parameters are estimated using individual data; not all data.
  - The estimate for K is questionable.
  - Ad hoc testing procedures $\alpha, Q_{max}, P_{max}$.
- Pros

# NLIN: Pros and Cons

- Cons
  - The parameters are estimated using individual data; not all data.
  - The estimate for K is questionable.
  - Ad hoc testing procedures $\alpha$, $Q_{max}$, $P_{max}$.
- Pros
  - Scientists are familiar with this approach

# Mixed-Effects Model

$$Y_{ij} = [C_0 + C_M I(Male_i = 1) + u_i] + e^{\log K}\{e^{-[(\alpha_0+\alpha_M I(Male_i=1)+b_i)P_j]} - 1\} + \epsilon_{ij}$$

$$\begin{pmatrix} u_i \\ b_i \end{pmatrix} \sim N_2(0, \Sigma)$$

$$\Sigma = \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_b \\ \rho\sigma_u\sigma_b & \sigma_b^2 \end{pmatrix}$$

- $C_M, \alpha_M$

# Mixed-Effects Model: Pros and Cons

- Pros

# Mixed-Effects Model: Pros and Cons

- Pros
  - The parameters are estimated using data from every individuals.

# Mixed-Effects Model: Pros and Cons

- Pros
  - The parameters are estimated using data from every individuals.
  - Hypothesis testings for gender effect come naturally within the model framework
- Cons

# Mixed-Effects Model: Pros and Cons

- Pros
  - The parameters are estimated using data from every individuals.
  - Hypothesis testings for gender effect come naturally within the model framework
- Cons
  - Scientists are not familiar with this approach

# Bayes' Theorem

$$f(\theta|data) = \frac{f(data|\theta)f(\theta)}{f(data)}$$

# Bayes' Theorem

$$f(\theta|data) = \frac{f(data|\theta)f(\theta)}{f(data)}$$

$$f(data) = \int f(data|\theta)f(\theta)d\theta$$

# Bayes' Theorem

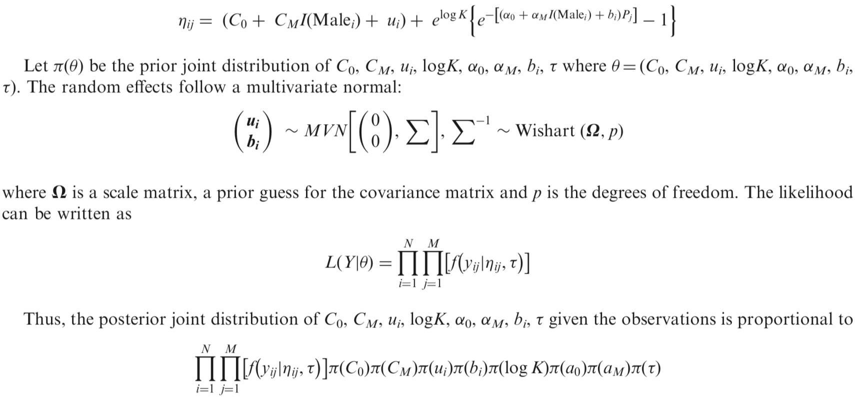
$$f(\theta|data) = \frac{f(data|\theta)f(\theta)}{f(data)}$$

$$f(data) = \int f(data|\theta)f(\theta)d\theta$$

$$f(\theta|data) \propto f(data|\theta)f(\theta)$$

A flate prior means all possible values are equally likely

# Bayesian Hierarchical Model Approach



$$\eta_{ij} = (C_0 + C_M I(\text{Male}_i) + u_i) + e^{\log K}\left\{e^{-[(\alpha_0 + \alpha_M I(\text{Male}_i) + b_i)P_j]} - 1\right\}$$

Let $\pi(\theta)$ be the prior joint distribution of $C_0$, $C_M$, $u_i$, $\log K$, $\alpha_0$, $\alpha_M$, $b_i$, $\tau$ where $\theta = (C_0, C_M, u_i, \log K, \alpha_0, \alpha_M, b_i, \tau)$. The random effects follow a multivariate normal:

$$\begin{pmatrix} \boldsymbol{u_i} \\ \boldsymbol{b_i} \end{pmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sum\right], \sum^{-1} \sim \text{Wishart}\ (\boldsymbol{\Omega}, p)$$

where $\boldsymbol{\Omega}$ is a scale matrix, a prior guess for the covariance matrix and $p$ is the degrees of freedom. The likelihood can be written as

$$L(Y|\theta) = \prod_{i=1}^{N}\prod_{j=1}^{M}[f(y_{ij}|\eta_{ij}, \tau)]$$

Thus, the posterior joint distribution of $C_0$, $C_M$, $u_i$, $\log K$, $\alpha_0$, $\alpha_M$, $b_i$, $\tau$ given the observations is proportional to

$$\prod_{i=1}^{N}\prod_{j=1}^{M}[f(y_{ij}|\eta_{ij}, \tau)]\pi(C_0)\pi(C_M)\pi(u_i)\pi(b_i)\pi(\log K)\pi(a_0)\pi(a_M)\pi(\tau)$$

# Comparison of the three approaches

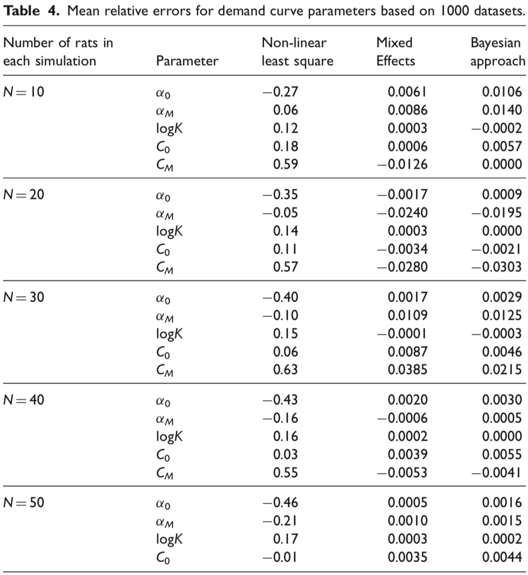


**Table 4.** Mean relative errors for demand curve parameters based on 1000 datasets.

| Number of rats in each simulation | Parameter | Non-linear least square | Mixed Effects | Bayesian approach |
|---|---|---|---|---|
| $N = 10$ | $\alpha_0$ | −0.27 | 0.0061 | 0.0106 |
| | $\alpha_M$ | 0.06 | 0.0086 | 0.0140 |
| | $\log K$ | 0.12 | 0.0003 | −0.0002 |
| | $C_0$ | 0.18 | 0.0006 | 0.0057 |
| | $C_M$ | 0.59 | −0.0126 | 0.0000 |
| $N = 20$ | $\alpha_0$ | −0.35 | −0.0017 | 0.0009 |
| | $\alpha_M$ | −0.05 | −0.0240 | −0.0195 |
| | $\log K$ | 0.14 | 0.0003 | 0.0000 |
| | $C_0$ | 0.11 | −0.0034 | −0.0021 |
| | $C_M$ | 0.57 | −0.0280 | −0.0303 |
| $N = 30$ | $\alpha_0$ | −0.40 | 0.0017 | 0.0029 |
| | $\alpha_M$ | −0.10 | 0.0109 | 0.0125 |
| | $\log K$ | 0.15 | −0.0001 | −0.0003 |
| | $C_0$ | 0.06 | 0.0087 | 0.0046 |
| | $C_M$ | 0.63 | 0.0385 | 0.0215 |
| $N = 40$ | $\alpha_0$ | −0.43 | 0.0020 | 0.0030 |
| | $\alpha_M$ | −0.16 | −0.0006 | 0.0005 |
| | $\log K$ | 0.16 | 0.0002 | 0.0000 |
| | $C_0$ | 0.03 | 0.0039 | 0.0055 |
| | $C_M$ | 0.55 | −0.0053 | −0.0041 |
| $N = 50$ | $\alpha_0$ | −0.46 | 0.0005 | 0.0016 |
| | $\alpha_M$ | −0.21 | 0.0010 | 0.0015 |
| | $\log K$ | 0.17 | 0.0003 | 0.0002 |
| | $C_0$ | −0.01 | 0.0035 | 0.0044 |

# Comparison of the three approaches

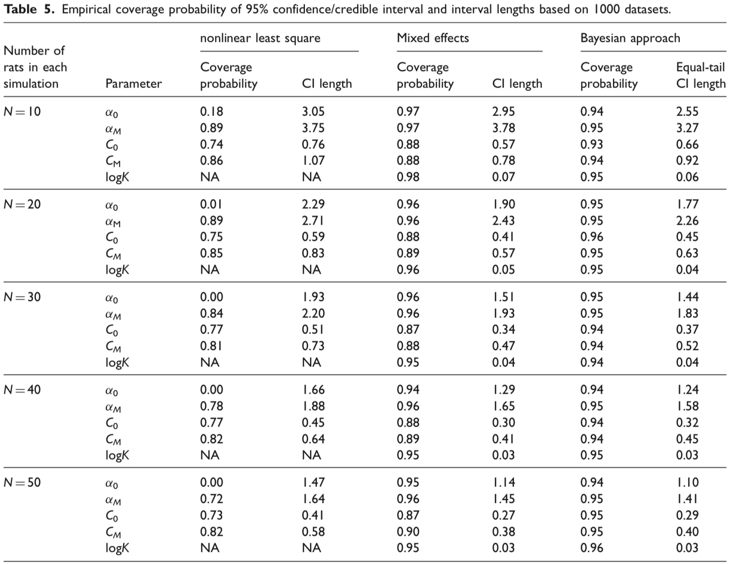


**Table 5.** Empirical coverage probability of 95% confidence/credible interval and interval lengths based on 1000 datasets.

| Number of rats in each simulation | Parameter | nonlinear least square | | Mixed effects | | Bayesian approach | |
|---|---|---|---|---|---|---|---|
| | | Coverage probability | CI length | Coverage probability | CI length | Coverage probability | Equal-tail CI length |
| $N = 10$ | $\alpha_0$ | 0.18 | 3.05 | 0.97 | 2.95 | 0.94 | 2.55 |
| | $\alpha_M$ | 0.89 | 3.75 | 0.97 | 3.78 | 0.95 | 3.27 |
| | $C_0$ | 0.74 | 0.76 | 0.88 | 0.57 | 0.93 | 0.66 |
| | $C_M$ | 0.86 | 1.07 | 0.88 | 0.78 | 0.94 | 0.92 |
| | $\log K$ | NA | NA | 0.98 | 0.07 | 0.95 | 0.06 |
| $N = 20$ | $\alpha_0$ | 0.01 | 2.29 | 0.96 | 1.90 | 0.95 | 1.77 |
| | $\alpha_M$ | 0.89 | 2.71 | 0.96 | 2.43 | 0.95 | 2.26 |
| | $C_0$ | 0.75 | 0.59 | 0.88 | 0.41 | 0.96 | 0.45 |
| | $C_M$ | 0.85 | 0.83 | 0.89 | 0.57 | 0.95 | 0.63 |
| | $\log K$ | NA | NA | 0.96 | 0.05 | 0.95 | 0.04 |
| $N = 30$ | $\alpha_0$ | 0.00 | 1.93 | 0.96 | 1.51 | 0.95 | 1.44 |
| | $\alpha_M$ | 0.84 | 2.20 | 0.96 | 1.93 | 0.95 | 1.83 |
| | $C_0$ | 0.77 | 0.51 | 0.87 | 0.34 | 0.94 | 0.37 |
| | $C_M$ | 0.81 | 0.73 | 0.88 | 0.47 | 0.94 | 0.52 |
| | $\log K$ | NA | NA | 0.95 | 0.04 | 0.94 | 0.04 |
| $N = 40$ | $\alpha_0$ | 0.00 | 1.66 | 0.94 | 1.29 | 0.94 | 1.24 |
| | $\alpha_M$ | 0.78 | 1.88 | 0.96 | 1.65 | 0.95 | 1.58 |
| | $C_0$ | 0.77 | 0.45 | 0.88 | 0.30 | 0.94 | 0.32 |
| | $C_M$ | 0.82 | 0.64 | 0.89 | 0.41 | 0.94 | 0.45 |
| | $\log K$ | NA | NA | 0.95 | 0.03 | 0.95 | 0.03 |
| $N = 50$ | $\alpha_0$ | 0.00 | 1.47 | 0.95 | 1.14 | 0.94 | 1.10 |
| | $\alpha_M$ | 0.72 | 1.64 | 0.96 | 1.45 | 0.95 | 1.41 |
| | $C_0$ | 0.73 | 0.41 | 0.87 | 0.27 | 0.95 | 0.29 |
| | $C_M$ | 0.82 | 0.58 | 0.90 | 0.38 | 0.95 | 0.40 |
| | $\log K$ | NA | NA | 0.95 | 0.03 | 0.96 | 0.03 |

# Nonparametric regression

Consider a continuous response with three predictors (although these methods can be extended to other types of response).

An additive model stipulates

$$Y_i = \mu + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + \epsilon_i,$$

and seeks to estimate the functions $f_1(x)$, $f_2(x)$, and $f_3(x)$ (typically via splines). These are fit in `proc gam` and `proc transreg`. We can also consider a transformation of $Y_i$ as well as pairwise interaction surfaces.

# Nonparametric regression

A parametric nonlinear model (Chapter 13) has a prespecified parametric form indexed by parameters $\gamma$

$$Y_i = f(\mathbf{x}_i, \gamma) + \epsilon_i.$$

For example the exponential growth/decay model is $Y_i = \gamma_0 e^{\gamma_1 x_i} + \epsilon_i$. Data reduction takes place through the estimation of $\gamma$ and $\sigma$.

Nonparametric regression is essentially unspecified

$$Y_i = f(\mathbf{x}_i) + \epsilon_i,$$

and seeks to estimate $f(\mathbf{x}) : \mathbb{R}^k \to \mathbb{R}$ directly. Two useful and popular methods are *lowess* and *kernel smoothing*.

# Kernel smoothing

Let's start with a univariate predictor yielding data $\{(x_i, Y_i)\}_{i=1}^n$. At each $x \in \mathbb{R}$, the kernel-smoothed estimate of $f(\cdot)$ is a weighted average of the $Y_i$'s:

$$\hat{f}_h(x) = \sum_{i=1}^n \left[ \frac{k\{(x_i - x)/h\}/h}{\sum_{j=1}^n k\{(x_j - x)/h\}/h} \right] Y_i.$$

Here, $k(d)$ is the kernel. Common choices are Gaussian $k(d) = e^{-0.5d^2}$ (most common), uniform $k(d) = I\{|d| < 1\}$, and Epanechnikov $k(d) = 0.75(1 - d^2)I\{|d| < 1\}$ (there are many more). Different kernel functions simply weight neighboring points differently.

# Bandwidth

The parameter $h$ is called the bandwidth. The larger the bandwidth, the smoother the estimate $\hat{f}_h$. What happens to $\hat{f}_h$ as $h \to \infty$? Is it possible to have $\hat{f}_h(x)$ outside the range of $Y_i$-values?

A common way to choose the bandwidth is through cross-validation, $\hat{h} = \text{argmin}_{h>0} \sum_{i=1}^{n} (Y_i - \hat{f}_{h,i}(x_i))^2$ where $\hat{f}_{h,i}$ is the kernel-smoothed estimate based on the $(n-1)$ pairs $\{(x_j, Y_j)\}_{j \neq i}$.

`ksmooth` in R gives kernel-smoothed regression estimates without standard errors. A great package that does a lot more (including handling categorical predictors) is `np`. You need to install it from CRAN.

# Yellowfin tuna example in R with Gaussian kernel-smoothing

Recall that $Y_i$ is length and $x_i$ is age. The default bandwidth $h$ selection is cross-validation; surprisingly, I found it under-smoothed the data. The default kernel is Gaussian.

```
library(np)
Tuna.df <- read.delim(‘‘Yellowfin.txt’’,header=T)
attach(Tuna.df)
Length_Pacific <- Length[Ocean==’’Pacific’’]
Age_Pacific <- Age[Ocean==’’Pacific’’]
fit1 <-npreg(Length_Pacific˜Age_Pacific)
plot(fit1,plot.errors.method="asymptotic",plot.errors.style="band",main="Kernel-smoothed")
points(Age_Pacific,Length_Pacific)
```

# 11.4 **LO**cally **WE**ighted **S**catterplot **S**moothing (lowess)

Kernel-smoothing is biased at the boundaries $\min\{x_i\}$ and $\max\{x_i\}$, and at the extrema of $f(\cdot)$. A method that solves some of these issues uses locally fitted polynomials to estimate $f(x)$ at each $x$ via weighted least squares (WLS). Lowess was introduced by Cleveland (1979).

Recall that weighted least squares weights some pairs $(x_i, Y_i)$ more heavily when "more information" is known about $Y_i$, e.g. $var(Y_i)$ is smaller than for other values. The weight $w_i$ attached to $(x_i, Y_i)$ is the $i$th diagonal of the matrix $\mathbf{W}$; the remaining elements are zero. The weighted least squares estimate of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (\mathbf{XWX}')^{-1}\mathbf{X}'\mathbf{WY}$.

# lowess

Consider estimating $f(x)$ at $x$ with a linear or quadratic function. If we assume that pairs $(x_i, Y_i)$ have more information for $f(x)$ at values of $x_i$ near $x$, we can weight them more using WLS. The most common weight function is tricube

$$w_i(x) = \left\{ \begin{array}{cc} [1-(|x-x_i|/d_q(x))^3]^3 & |x-x_i| < d_q(x) \\ 0 & |x-x_i| > d_q(x) \end{array} \right\}.$$

$d_q(x)$ is a distance such that the proportion of $x_i$ values within $x$ is $q$, i.e. $d_q(x) = \min\{d > 0 : \frac{1}{n}\sum_{i=1}^{n} I\{|x_i - x| < d\} \geq q\}$. A common choice of $q$ is 0.5 (p. 450).

The lowess estimate of $f(x)$, assuming local linear fitting, is then $\hat{f}(x) = [\begin{array}{cc} 1 & x \end{array}](\mathbf{X}\mathbf{W}(x)\mathbf{X}')^{-1}\mathbf{X}'\mathbf{W}(x)\mathbf{Y}$ where $\mathbf{W}(x) = \text{diag}(w_1(x), \dots, w_n(x))$ and the $i$th row of $bX$ is $[\begin{array}{cc} 1 & x_i \end{array}]$. For *each value* of $x$, a separate WLS is fitted – lowess requires *a lot* of computation!

# Yellowfin tuna example in R with lowess

This uses defaults, which actually over-smooth in this case (enp.target can be manipulated to fix this). An older function is lowess; loess has improvements on lowess but gives essentially the same answers.

```
fit2=loess(Length_Pacific~Age_Pacific)
pred.Age=seq(0,1200,20)
pred2=predict(fit2,pred.Age,se=TRUE)
plot(pred.Age,pred2$fit,type="l",xlab="Age",ylab="Length",main="Lowess Fit")
lines(pred.Age,pred2$fit-1.96*pred2$se.fit,lty=3)
lines(pred.Age,pred2$fit+1.96*pred2$se.fit,lty=3)
points(Age_Pacific,Length_Pacific)
```

# Similarities between lowess and kernel-smoothing

Both kernel-smoothing and lowess have weight functions and bandwidths that determine how points in a neighborhood of $x$ are weighted.

Both estimates are written as $\hat{f}(x) = \mathbf{c}(x)'\mathbf{Y}$, i.e. are linear combinations of the $Y_i$'s that depend on $\mathbf{x}$. In STAT 704 regression, $\hat{f}(x) = \mathbf{c}(x)'\mathbf{Y}$ where $\mathbf{c}(x)' = [\begin{array}{cc} 1 & x \end{array}](\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Note that kernel-smoothing provides a true average of the $Y_i$'s at each point, whereas lowess values of $c_i(x)$ may be negative or greater than one.

Both methods are generalized to more than one predictor similarly. Predictors are standardized to have variance one and Euclidean distance $d = ||\mathbf{x} - \mathbf{x}^*||$ is used in the weight function rather than $|x - x^*|$, or else the Mahalanobis distance is used $d = \sqrt{(\mathbf{x} - \mathbf{x}^*)'\mathbf{S}^{-1}(\mathbf{x} - \mathbf{x}^*)}$ (no need to standardized first). Note that categorical predictors need some thought.

# Questions and comments

- Is extrapolation a good idea with lowess or kernel-smoothed methods?
- The asymptotics for nonparametric smoothing methods is worth an entire course. A bit is covered in STAT 824 (nonparametrics).
- Which method, lowess or kernel-smoothing, is more appropriate for Bernoulli data? Why?
- There's some nice animation here: http://www.r-bloggers.com/some-heuristics-about-local-regression-and-kernel-smoothing/
- A method worthy of its own lecture is *basis expansions*. Basis expansions write the unknown $f(\cdot)$ as $f(\mathbf{x}) = \sum_{k=1}^{K} \beta_k \phi_k(\mathbf{x})$ for a set of known functions $\phi_k(\cdot)$. The unknown parameters are $\beta_1, \ldots, \beta_K$. *This yields a linear model.*
- Example basis expansions include polynomials, Legendre polynomials, wavelets, sines and cosines, and B-splines.