应用回归分析

上海财经大学　统计与管理学院

第六章多项式回归与因子回归

章节概括：

- 多项式回归
- Delta方法
- 线性组合与过参数化
- 单因子(因素)回归
- 多因子回归
- POD模型

多项式回归

- 一个自变量
- 线性回归

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$$

- 多项式回归

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$$

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d$$

二项回归曲线

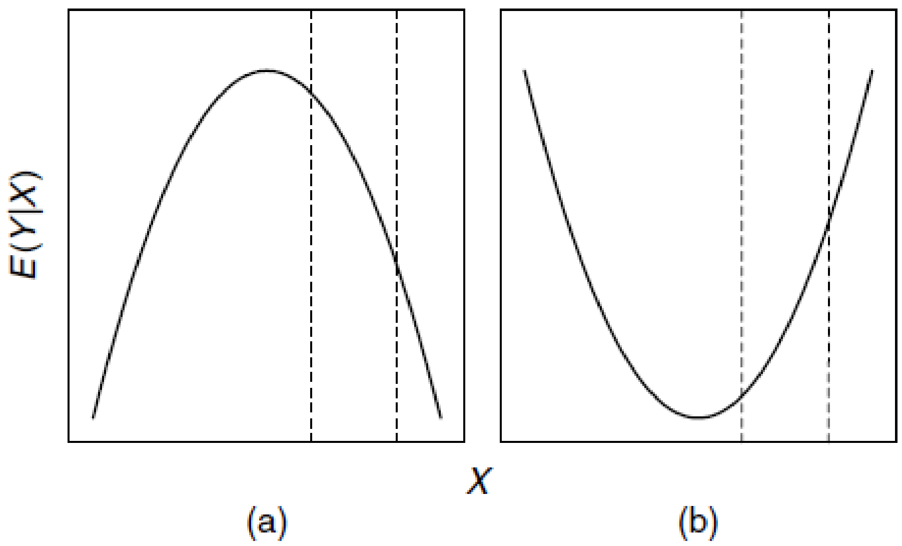

FIG. 6.1 Generic quadratic curves. A quadratic is the simplest curve that can approximate a mean function with a minimum or maximum within the range of possible values of the predictor. It can also be used to approximate some nonlinear functions without a minimum or maximum in the range of interest, possibly using the part of the curve between the dashed lines.

多项式回归

- 一个变多个(d+1)个

$$X, X^2, \ldots, X^d$$

- 模型 $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$

- 估计 $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

多项式回归

- 二个自变量

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$

一次项、二次向、交叉项

- 2个变（2＋2＋1)个
- k个变 k+k+k(k+1)/2个

多项式回归

$$\mathrm{E}(Y|X_1 = x_1 + \delta, X_2 = x_2) = \beta_0 + \beta_1(x_1 + \delta) + \beta_2 x_2 + \beta_{11}(x_1 + \delta)^2 + \beta_{22}x_2^2 + \beta_{12}(x_1 + \delta)x_2$$

$$\mathrm{E}(Y|X_1 = x_1 + \delta, X_2 = x_2) - \mathrm{E}(Y|X_1 = x_1, X_2 = x_2) = (\beta_{11}\delta^2 + \beta_1\delta) + 2\beta_{11}\delta x_1 + \beta_{12}\delta x_2$$

例子

$$E(Y|X_1, X_2) = -2204.4850 + 25.9176X_1 + 9.9183X_2$$
$$-0.1569X_1^2 - 0.0120X_2^2 - 0.0416X_1X_2$$

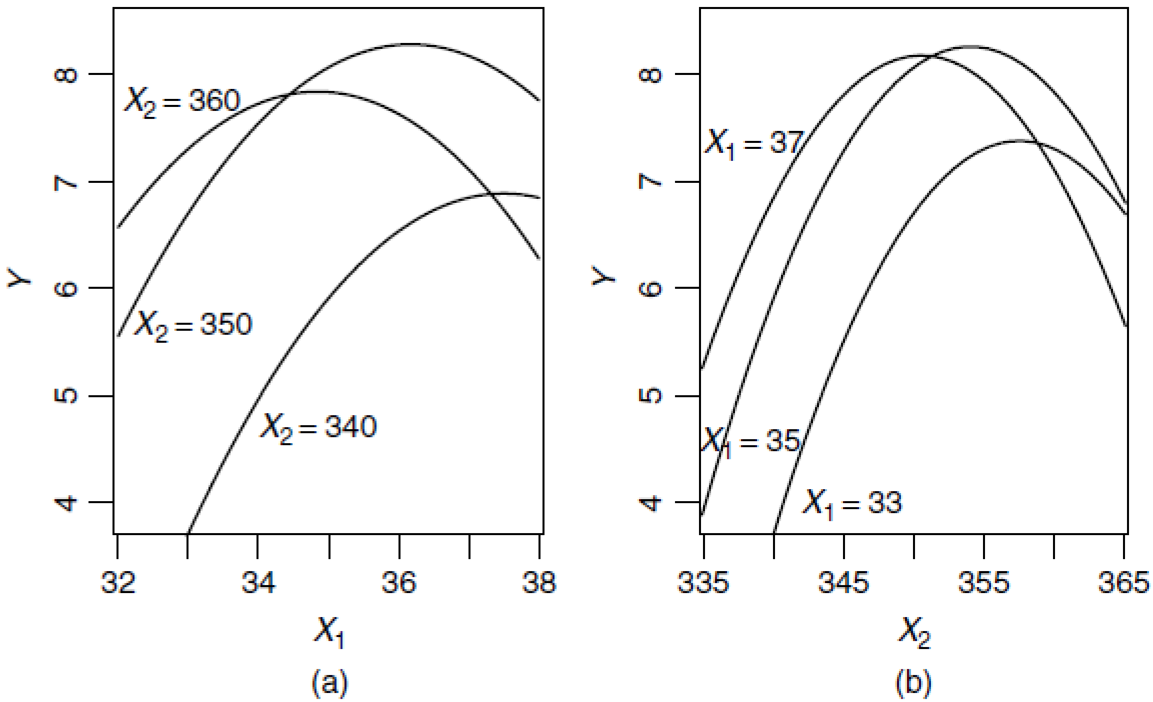

FIG. 6.3 Estimated response curves for the cakes data, based on (6.7).

例子

- $\beta_{12} = 0$

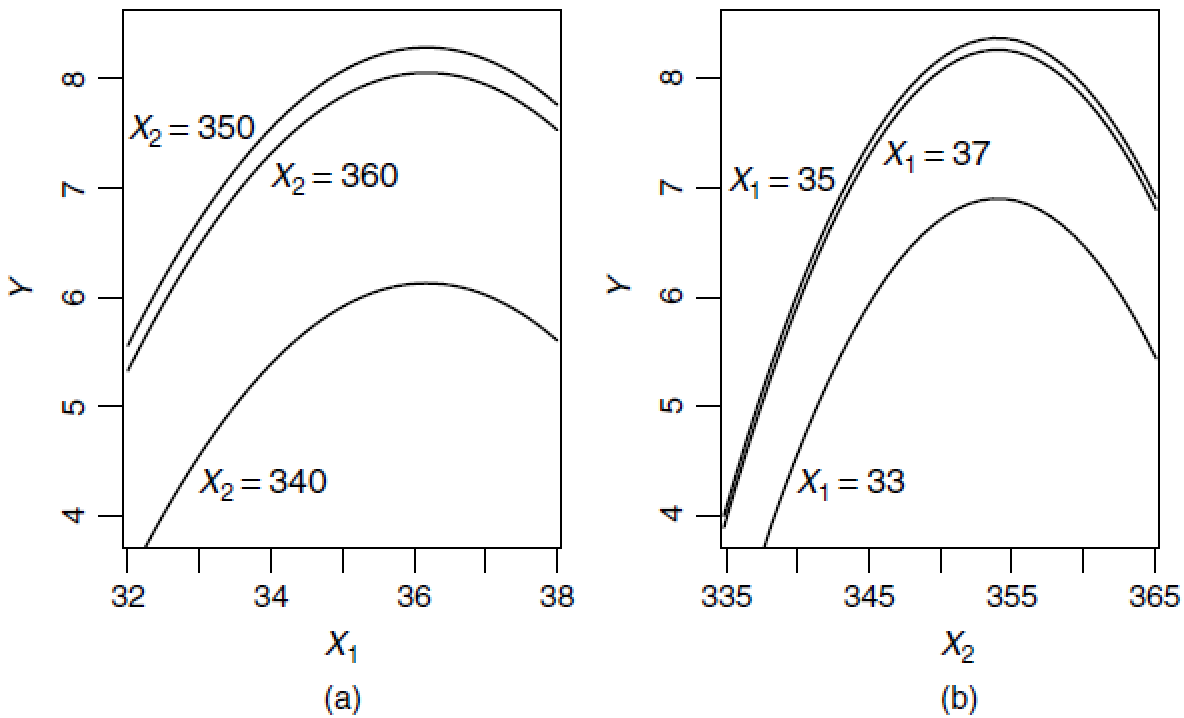

FIG. 6.4 Estimated response curves for the cakes data, based on fitting with $\beta_{12} = 0$.

Delta方法

- 假设 $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \sigma^2 \mathbf{D})$

  $g(\boldsymbol{\theta})$ 连续非线性

- Taylor 展开

$$g(\hat{\boldsymbol{\theta}}) = g(\boldsymbol{\theta}^*) + \sum_{j=1}^{k} \frac{\partial g}{\partial \theta_j}(\hat{\theta}_j - \theta_j^*) + \text{small terms}$$

$$\approx g(\boldsymbol{\theta}^*) + \dot{\mathbf{g}}(\boldsymbol{\theta}^*)'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$$

$$\dot{\mathbf{g}}(\boldsymbol{\theta}^*) = \frac{\partial g}{\partial \boldsymbol{\theta}} = \left(\frac{\partial g}{\partial \theta_1}, \ldots, \frac{\partial g}{\partial \theta_k}\right)'$$

Delta方法

- $$\begin{aligned}\mathrm{Var}(g(\hat{\theta})) &= \mathrm{Var}(g(\theta^*)) + \mathrm{Var}\left[\dot{\mathbf{g}}(\theta^*)'(\hat{\theta} - \theta^*)\right] \\ &= \dot{\mathbf{g}}(\theta^*)'\mathrm{Var}(\hat{\theta})\dot{\mathbf{g}}(\theta^*) \\ &= \sigma^2\dot{\mathbf{g}}(\theta^*)'\mathbf{D}\dot{\mathbf{g}}(\theta^*)\end{aligned}$$

$$\mathrm{Var}(g(\hat{\theta})) = \sigma^2 \sum_{i=1}^{k}\sum_{j=1}^{k} \dot{g}_i \dot{g}_j d_{ij}$$

例子

- $E(Y|X=x)=\beta_0+\beta_1 x+\beta_2 x^2$
- 最大/小值 $dE(Y|X=x)/dx=0,$

$$x_M=-\beta_1/(2\beta_2)$$

- $g(\beta)=\ -\beta_1/(2\beta_2)$
- $$\left(\frac{\partial g}{\partial \beta}\right)'=\left(0,-\frac{1}{2\hat{\beta}_2},\frac{\hat{\beta}_1}{2\hat{\beta}_2^2}\right)$$

$$\mathrm{Var}(g(\beta))=\frac{1}{4\hat{\beta}_2^2}\left(\mathrm{Var}(\hat{\beta}_1)+\frac{\hat{\beta}_1^2}{\hat{\beta}_2^2}\mathrm{Var}(\hat{\beta}_2)-\frac{2\hat{\beta}_1}{\hat{\beta}_2}\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2)\right)$$

线性组合

- 例子：Berkeley Guidance Study
- N=70 (girls only)
- Y = soma
- X = WT2, WT9, WT18

散点图

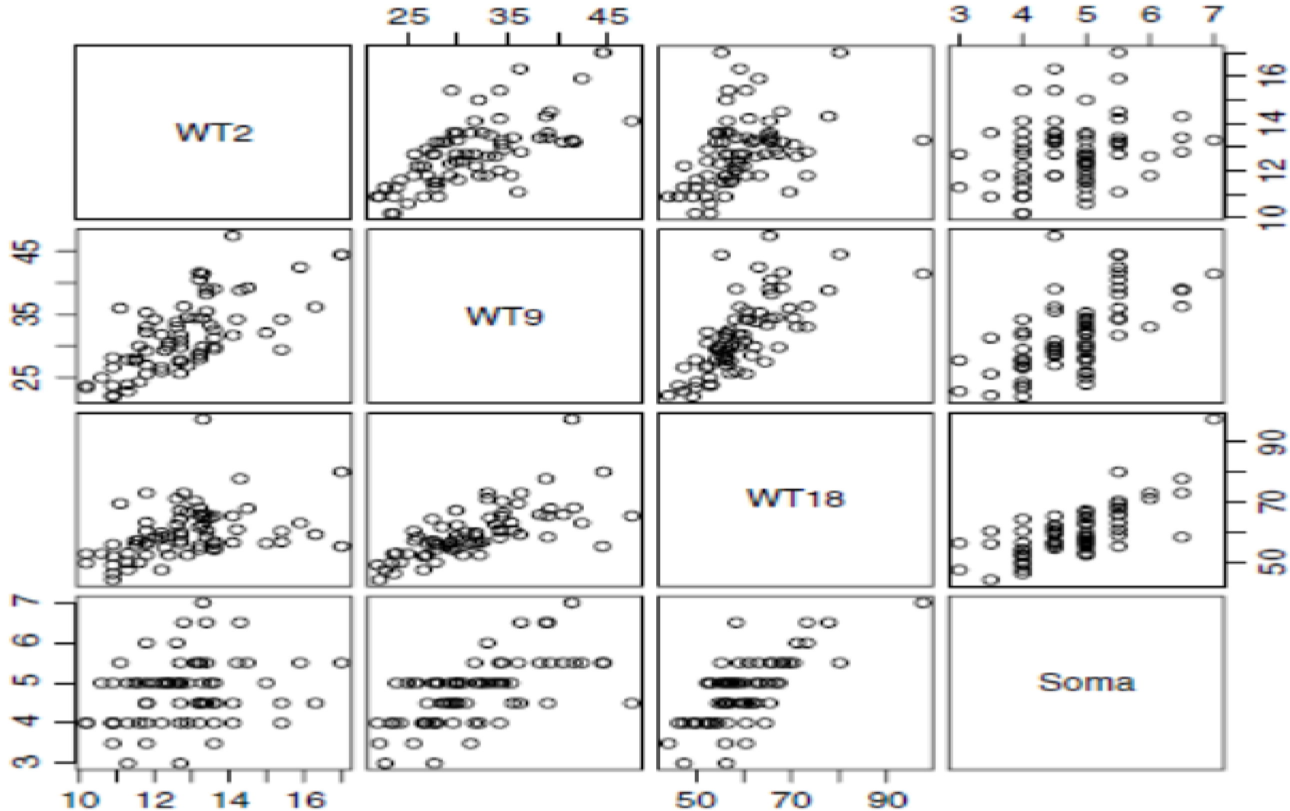

FIG. 4.1 Scatterplot matrix for the girls in the Berkeley Guidance Study.

线性组合

- Y ~ WT2 + WT9 + WT18
- Y ~ WT2 + DW9 + DW18

$$\begin{aligned} WT2 &= \text{Weight at age 2} \\ DW9 = WT9 - WT2 &= \text{Weight gain from age 2 to 9} \\ DW18 = WT18 - WT9 &= \text{Weight gain from age 9 to 18} \end{aligned}$$

- Y ~ AVE + LIN + QUAD

$$AVE = (WT2 + WT9 + WT18)/3$$

$$LIN = WT18 - WT2$$

$$QUAD = WT2 - 2WT9 + WT18$$

模型对比

TABLE 4.1 Regression of *Soma* on Different Combinations of Three Weight Variables for the $n = 70$ Girls in the Berkeley Guidance Study

Term	Model 1	Model 2	Model 3
(Intercept)	1.5921	1.5921	1.5921
WT2	–0.1156	–0.0111	–0.1156
WT9	0.0562		0.0562
WT18	0.0483		0.0483
DW9		0.1046	NA
DW18		0.0483	NA

过参数化

- Y ~ WT2 + WT9 + WT18 + DW9 + DW18
- 设计矩阵$\boldsymbol{X}$中的自变量列之间不相关, **X**是一满秩矩阵
- 线性强相关 / 岭回归

因子回归

- 非度量型（质的）因素自变量

  性别：男/女；颜色：红黄蓝…

- 两个水平： 0/1, -1/1, … ???
- 多个水平： (1,2,3,4,5), (2, 3, 5, 7, 11) ???

单个因子

- D个水平
- 第j个因子 U_j
- 第i个观测的第j个因子 (虚拟变量dummy variable)

$$u_{ij} = \begin{cases} 1 & \text{if } D_i = j\text{th category of } D \\ 0 & \text{otherwise} \end{cases}$$

-

U_1	U_2	U_3
1	0	0
1	0	0
0	0	1
0	1	0
0	1	0
0	0	1
1	0	0

睡眠数据

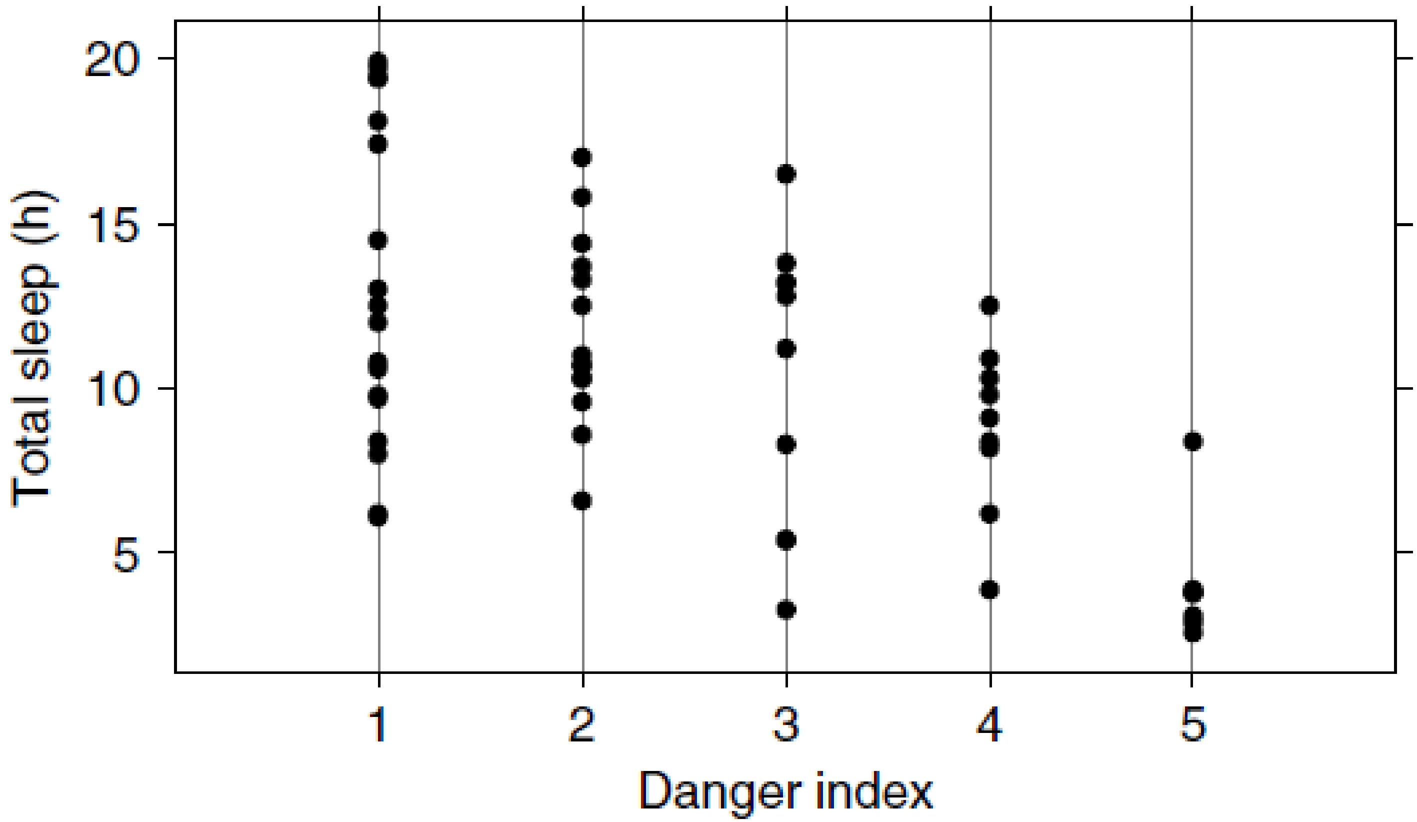

FIG. 6.5 Total sleep versus danger index for the sleep data.

单因子回归

- 无截距项

$$E(TS|D) = \beta_1 U_1 + \beta_2 U_2 + \beta_3 U_3 + \beta_4 U_4 + \beta_5 U_5$$

- 有截距项

$$E(TS|D) = \eta_0 + \eta_2 U_2 + \eta_3 U_3 + \eta_4 U_4 + \eta_5 U_5$$

- d个水平， 1 个限制，(d-1)个虚拟变量

- 参数解释

单因子方差分析

	Estimate	Std. Error	t-value	Pr(>|t|)
(a) Mean function (6.15)				
U_1	13.0833	0.8881	14.73	0.0000
U_2	11.7500	1.0070	11.67	0.0000
U_3	10.3100	1.1915	8.65	0.0000
U_4	8.8111	1.2559	7.02	0.0000
U_5	4.0714	1.4241	2.86	0.0061

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
D	5	6891.72	1378.34	97.09	0.0000
Residuals	53	752.41	14.20		

单因子方差分析

	Estimate	Std. Error	t-value	Pr(>|t|)
(b) Mean function (6.16)				
Intercept	13.0833	0.8881	14.73	0.0000
U_2	−1.3333	1.3427	−0.99	0.3252
U_3	−2.7733	1.4860	−1.87	0.0675
U_4	−4.2722	1.5382	−2.78	0.0076
U_5	−9.0119	1.6783	−5.37	0.0000

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
D	4	457.26	114.31	8.05	0.0000
Residuals	53	752.41	14.20		

四个模型

- 一个因子+一个自变量
- 5个水平的危险因子D
- 数值型的自变量log(BodyWt)

$$\mathrm{E}(TS|\log(BodyWt) = x, D = j) = \beta_{0j} + \beta_{1j}x$$

模型1

- 最广义的
- 每个水平都有对应的截距和斜率
- 表示I

$$\mathrm{E}(TS|\log(BodyWt) = x, D = j) = \sum_{j=1}^{d} (\beta_{0j} U_j + \beta_{1j} U_j x)$$

- 2d个参数
- R

$$TS \sim -1 + D + D{:}\log(BodyWt)$$

模型1

- 表示II

$$\mathrm{E}(TS|\log(BodyWt)=x, D=j)=\eta_0+\eta_1 x+\sum_{j=2}^{d}(\eta_{0j}U_j+\eta_{1j}U_j x)$$

- 参数解释 $\eta_0=\beta_{01},\ \eta_1=\beta_{11}$

$$j>1,\ \eta_{0j}=\beta_{0j}-\beta_{01} \text{ and } \eta_{1j}=\beta_{1j}-\beta_{11}$$

- R

$$\log(TS)\sim\log(BodyWt)+D+D{:}\log(BodyWt)$$

模型2

- 平行回归

$$\beta_{11} = \beta_{12} = \cdots = \beta_{1d}$$

$$\eta_{12} = \eta_{12} = \cdots = \eta_{1d} = 0$$

- 无交叉项, 同斜率
- (d+1)参数
- R

$$\log(TS) \sim D + \log(BodyWt)$$

模型3

- 同截距

$$\beta_{01} = \cdots = \beta_{0d}$$

$$\eta_{02} = \cdots = \eta_{0d} = 0$$

- (d+1)参数
- R

$$TS \sim 1 + D{:}\log(BodyWt)$$

模型4

- 所有水平的截距和斜率都相同

$$\beta_{01} = \cdots = \beta_{0m} \qquad \beta_{11} = \cdots = \beta_{1m}$$

$$\eta_{02} = \cdots = \eta_{0d} = \eta_{12} = \cdots = \eta_{1d} = 0$$

- 2个参数
- R

$$TS \sim \log(BodyWt)$$

四个模型

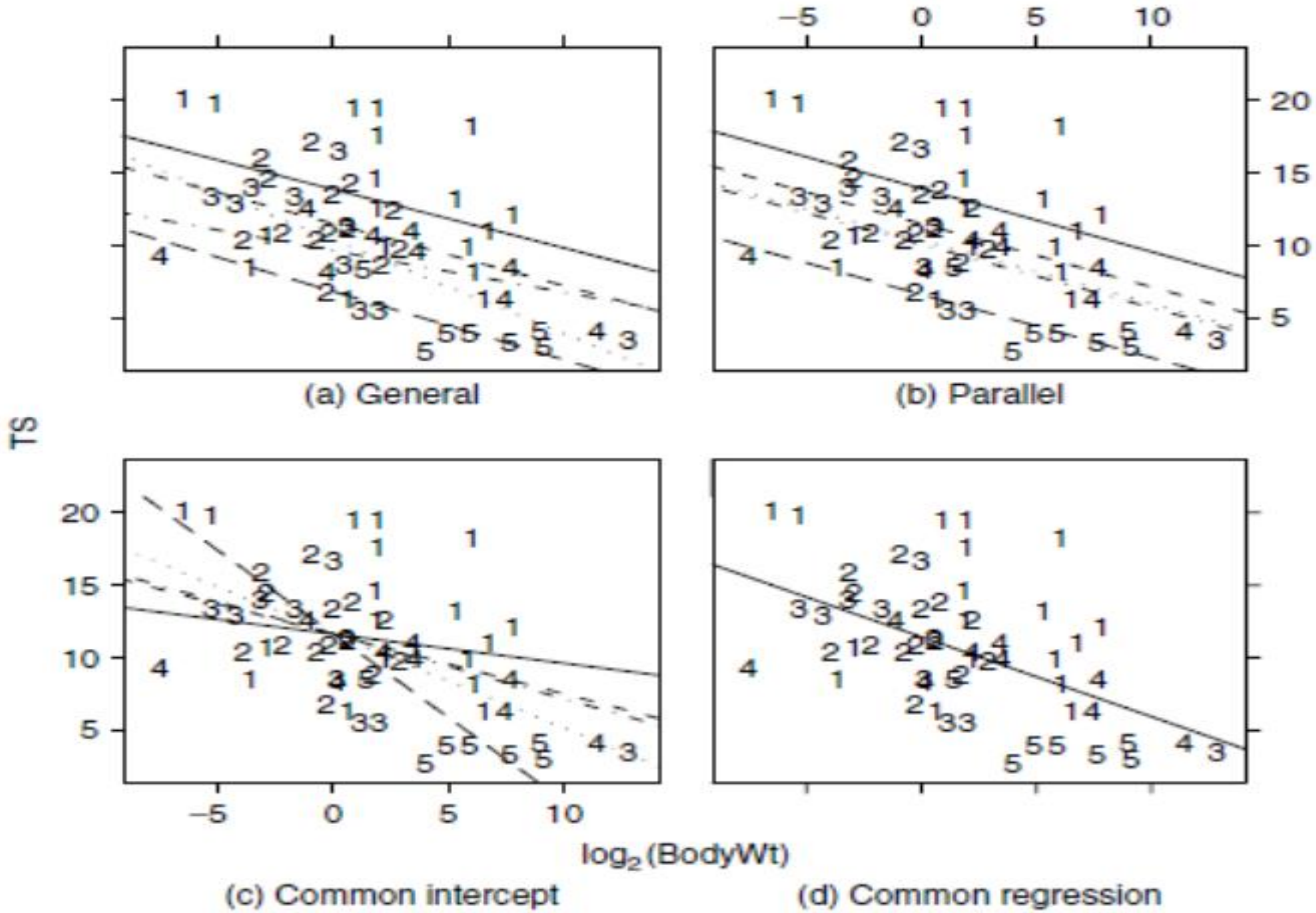

FIG. 6.6 Four models for the regression of *TS* on log(*BodyWt*) with five groups determined by *D*.

四个模型

- F检验

$$F_\ell = \frac{(RSS_\ell - RSS_1)(df_\ell - df_1)}{RSS_1/df_1}$$

TABLE 6.2 Residual Sum of Squares and df for the Four Mean Functions for the Sleep Data

	df	RSS	F	P(>F)
Model 1, most general	48	565.46		
Model 2, parallel	52	581.22	0.33	0.853
Model 3, common intercept	52	709.49	3.06	0.025
Model 4, all the same	56	866.23	3.19	0.006

多因子回归

- wool数据
- 3个因子，每个因子有3个水平
- 3^3 = 27 组合

TABLE 6.3 The Wool Data

Variable	Definition
Len	Length of test specimen (250, 300, 350 mm)
Amp	Amplitude of loading cycle (8, 9, 10 mm)
Load	Load put on the specimen (40, 45, 50 g)
log(*Cycles*)	Logarithm of the number of cycles until the specimen fails

多因子回归

- 截距 + 每个因子2个虚拟变量x 3个因子＝ 7
- (双因子交叉项2x2) x 3 = 12
- 三因子交叉项　2x2x2 = 8
- 7+12+8 = 27

$$\log(Cycles) \sim Len + Amp + Load$$

$$\log(Cycles) \sim Len + Amp + Load + Len{:}Amp + Len{:}Load + Amp{:}Load$$

$$\log(Cycles) \sim Len + Amp + Load + Len{:}Amp + Len{:}Load + Amp{:}Load + Len{:}Amp{:}Load$$

POD模型

- 自变量 $X = (X_1, \ldots, X_p)$
- 因子 F

- 同斜率、无交叉项
- $$Y \sim 1 + F + X_1$$
  $$Y \sim 1 + F + X_2$$
  $$Y \sim 1 + F + X_1 + X_2$$
  $$Y \sim 1 + F + X_1 + X_2 + X_1X_2$$

POD模型

- Partial One-Dimensional

$$\mathrm{E}(Y|X = \mathrm{x}, F = j) = \eta_{0j} + \eta_{1j}(\mathrm{x}'\beta^*)$$

- 最广义模型
- 非线性模型，最小二乘法失效

例子

- Australian Athletes
- N = 202
- Y = LBM (lean body mass)
- X = Sex, Ht, Wt, RCC (red cell count)

散点图

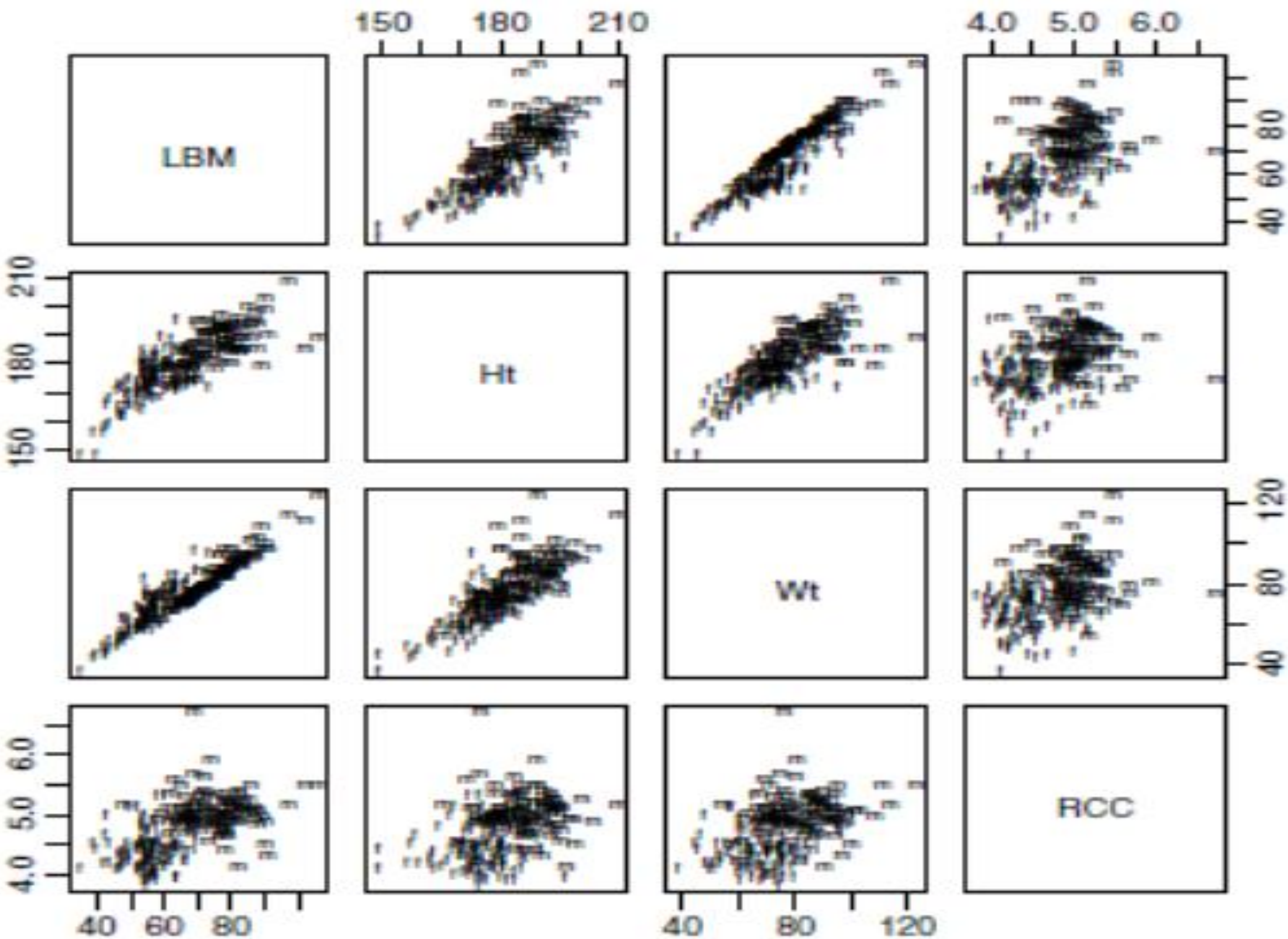

FIG. 6.7 Scatterplot matrix for the Australian athletes data, using "m" for males and "f" for females.

模型对比

$$E(LBM|Sex, Ht, Wt, RCC) = \beta_0 + \beta_1 Sex + \beta_2 Ht + \beta_3 Wt + \beta_4 RCC$$

$$E(LBM|Sex, Ht, Wt, RCC) = \beta_0 + \beta_1 Sex + \beta_2 Ht + \beta_3 Wt + \beta_4 RCC + \beta_{12}(Sex \times Ht) + \beta_{13}(Sex \times Wt) + \beta_{14}(Sex \times RCC) \quad (6.25)$$

$$E(LBM|Sex, Ht, Wt, RCC) = \beta_0 + \beta_1 Sex + \beta_2 Ht + \beta_3 Wt + \beta_4 RCC + \eta_0 Sex + \eta_1 Sex \times (\beta_2 Ht + \beta_3 Wt + \beta_4 RCC)$$

POD模型

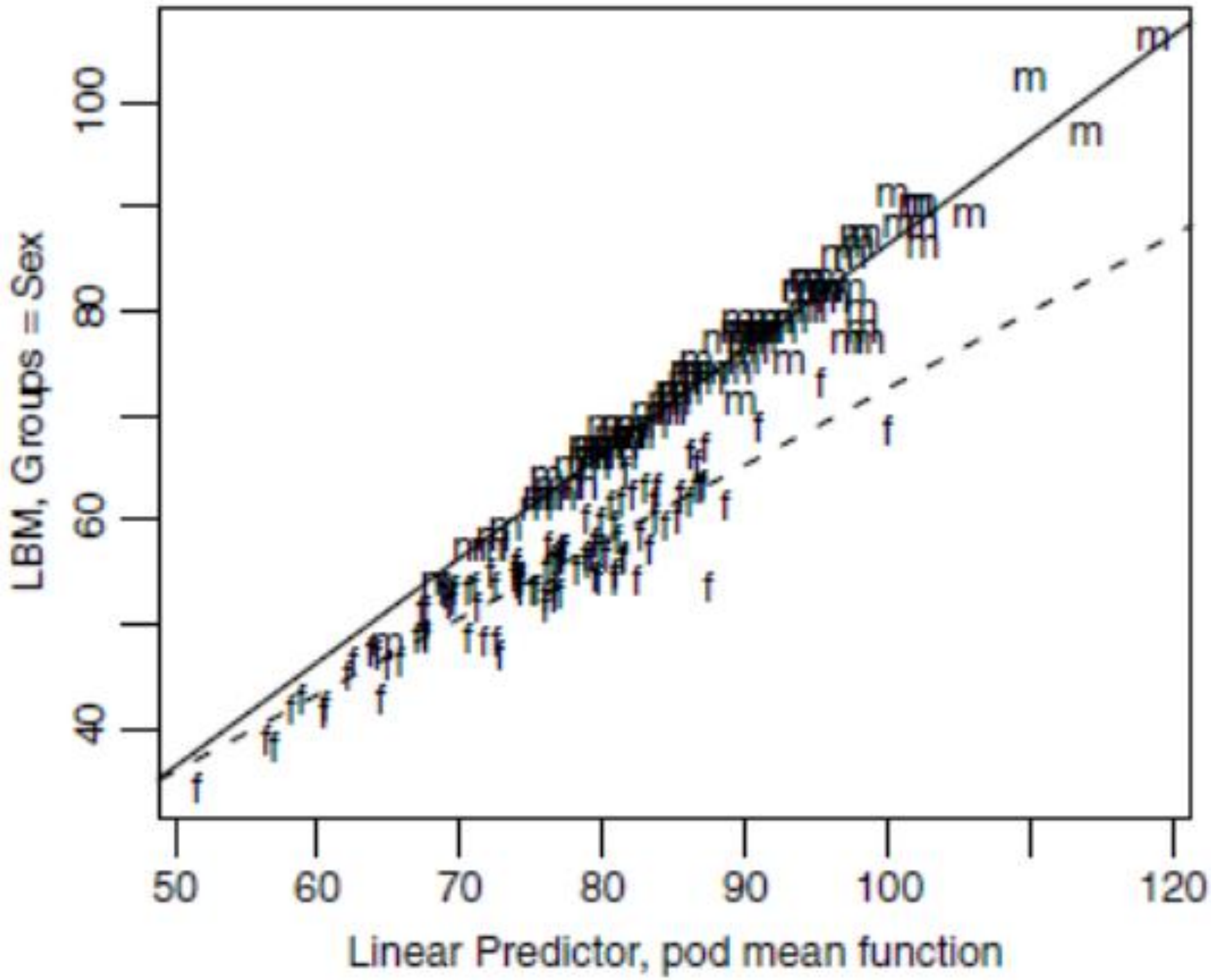

FIG. 6.8 Summary graph for the POD mean function for the Australian athletes data.

Thank You !