STA302/1001: Methods of Data Analysis

Instructor: Fang Yao

Chapter 6: Polynomials and Factors

Polynomials

- what shall we do if lack of fit exists?
- we could do nothing and just sit there and cry
- or we could improve our model
- Polynomial Regression: some terms are higher power of some predictors
- simplest example: quadratic regression

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

- a natural question: use straight line or quadratic?

Polynomials - con't

- answer by F-test from multiple regression ANOVA
- in general:

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

- important question: how to choose d
- e.g. find the most desirable value of x that maximizes or minimizes $E(Y|X)$ in quadratic regression
- for $\mathrm{E}(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$, solving

$$\frac{d\mathrm{E}(Y|X=x)}{dx} = 0 \quad \Rightarrow \quad x_M = \frac{-\beta_1}{2\beta_2}$$

Polynomials with Several Predictors

- a special case of two predictors:

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

- the term $X_1 X_2$ is called an <u>interaction</u>
- effect of X_2 cannot be kept constant if we change X_1
- if we only limit the highest order to 2, how many terms are there for k predictors?
- one intercept, k linear terms, k quadratic terms and $\frac{k(k-1)}{2}$ interaction terms
- e.g., $k = 5$, altogether 21 terms

Polynomials with Several Predictors - con't

- Y: palatability score; X_1: baking time; X_2: baking temp

with interaction

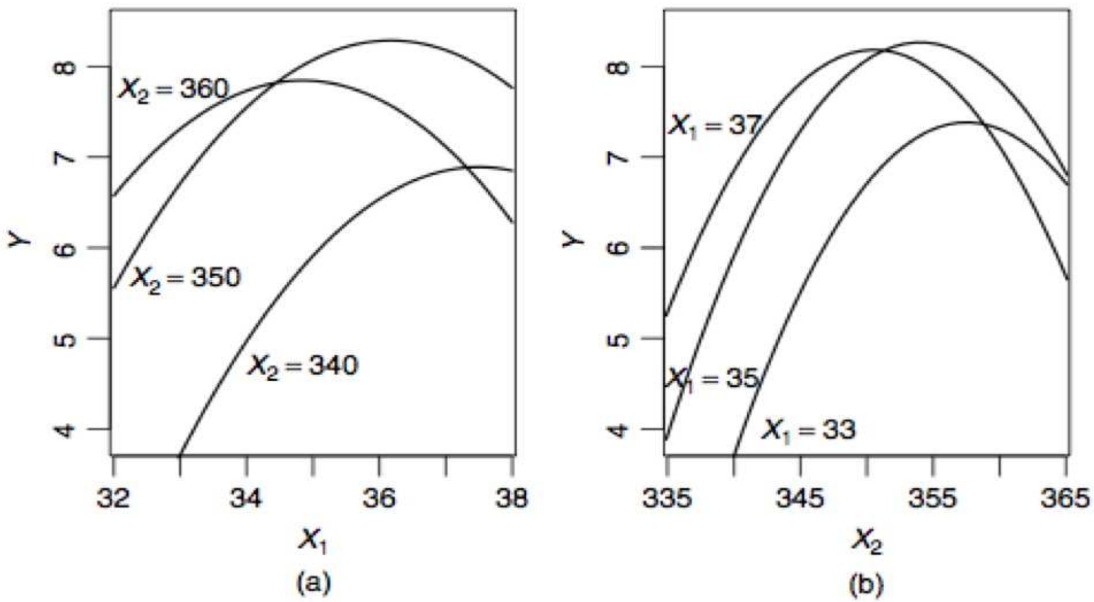


FIG. 6.3 Estimated response curves for the cakes data, based on (6.7).

Polynomials with Several Predictors - con't

without interaction

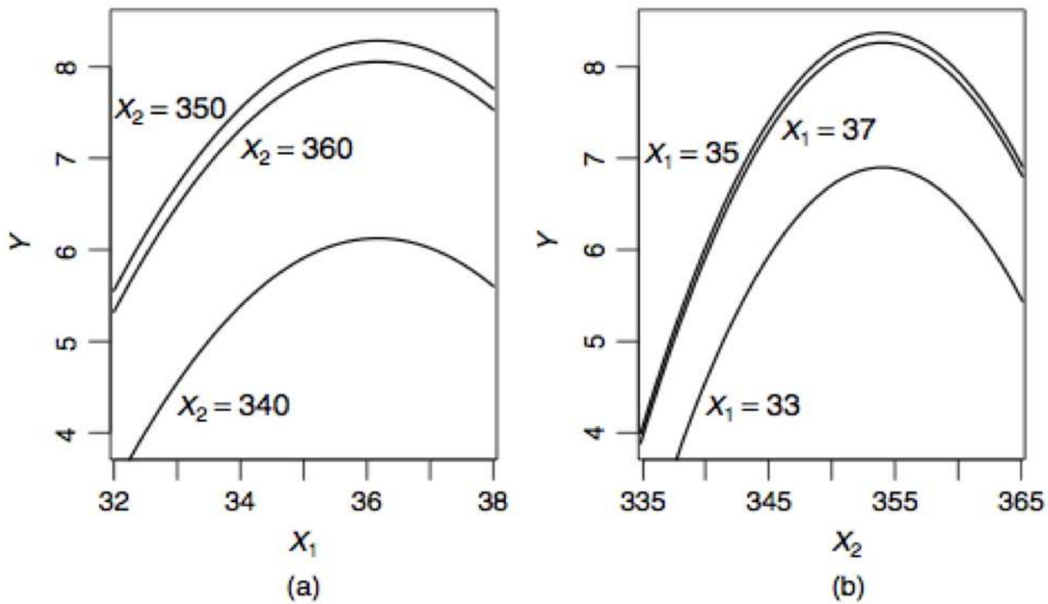


FIG. 6.4 Estimated response curves for the cakes data, based on fitting with $\beta_{12} = 0$.

The Delta Method

- provides approximate standard errors for nonlinear combinations of parameter estimates
- e.g., what is $\mathrm{Var}(\hat{x}_M)$ where $\hat{x}_M = \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$?
- suppose $\hat{\boldsymbol{\theta}} \overset{\circ}{\sim} N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and $g(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$ ($\boldsymbol{\theta}$ may be a vector)
- then, when n is large, we have

$$\begin{aligned}
\mathrm{E}[g(\hat{\boldsymbol{\theta}})] &\approx g(\boldsymbol{\theta}) \\
\mathrm{Var}[g(\hat{\boldsymbol{\theta}})] &\approx \dot{g}(\boldsymbol{\theta})' \boldsymbol{\Sigma} \dot{\mathbf{g}}(\boldsymbol{\theta}) \\
\text{where } \dot{g}(\boldsymbol{\theta}) &= \frac{\partial g}{\partial \boldsymbol{\theta}} = \left(\frac{\partial g}{\partial \theta_1}, \cdots, \frac{\partial g}{\partial \theta_k}\right)'
\end{aligned}$$

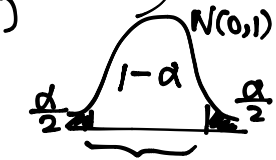
- note: some authors use $\sigma^2 \mathrm{D}$ instead of $\boldsymbol{\Sigma}$

The Delta Method - con't

$$\frac{g(\hat{\beta})-g(\beta)}{\sqrt{\dot{g}(\beta)'\Sigma\dot{g}(\hat{\beta})}} \dot{\sim} N(0,1)$$

- back to the example for $\hat{x}_M$
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)'$ $\quad g(\hat{\beta}) \dot{\sim} N(g(\beta), \dot{g}(\beta)\Sigma \dot{g}(\beta))$
- we know, for large n, $\hat{\boldsymbol{\beta}} \overset{\circ}{\sim} N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
- R function vcov(lm.fit) gives $\widehat{\text{Cov}}(\hat{\beta}) \approx \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$
- $g(\hat{\boldsymbol{\beta}}) = \frac{-\hat{\beta}_1}{2\hat{\beta}_2}. \Rightarrow \dot{g}(\hat{\boldsymbol{\beta}}) = (0, \frac{-1}{2\hat{\beta}_2}, \frac{\hat{\beta}_1}{2\hat{\beta}_2^2})$

$$P\left(\frac{|g(\hat{\beta})-g(\beta)|}{\sqrt{\dot{g}(\hat{\beta})'\hat{\Sigma}\dot{g}(\hat{\beta})}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$



$$\begin{aligned} \text{Var}(g(\hat{\boldsymbol{\beta}})) &= \dot{g}(\hat{\boldsymbol{\beta}})'\widehat{\text{Cov}}(\hat{\beta})\dot{g}(\hat{\boldsymbol{\beta}}) \\ &= \frac{1}{4\hat{\beta}_2^2}\Big(\text{Var}(\hat{\beta}_1) + \frac{\hat{\beta}_1^2}{\hat{\beta}_2^2}\text{Var}(\hat{\beta}_2) - \frac{2\hat{\beta}_1}{\hat{\beta}_2}\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)\Big) \end{aligned}$$

- use z-test or z-interval, i.e., critical value from $N(0,1)$

The Delta Method - con't

- revisit cakes data: find optimal baking times given different baking temperatures
- x_1: baking time; x_2: baking temperature
  $\mathrm{E}(Y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$
- solve for optimal baking time: $x_M = g(\boldsymbol{\beta}; x_2) = -\frac{\beta_1 + \beta_5 x_2}{2\beta_3}$
- $\frac{\partial x_M}{\partial \boldsymbol{\beta}} = \dot{g}(\boldsymbol{\beta}; x_2) = (0, -\frac{1}{2\beta_3}, 0, \frac{\beta_1 + \beta_5 x_2}{2\beta_3^2}, 0, -\frac{x_2}{2\beta_3})'$
- $\mathrm{Var}(\hat{x}_M) = \dot{g}(\hat{\boldsymbol{\beta}}; x_2)' \widehat{\mathrm{Cov}}(\hat{\beta}) \dot{g}(\hat{\boldsymbol{\beta}}; x_2)$
- $100(1-\alpha)$% pointwise confidence interval for x_M:

$$\hat{x}_M \pm z_{\alpha/2} \sqrt{\dot{g}(\hat{\boldsymbol{\beta}}; x_2)' \widehat{\mathrm{Cov}}(\hat{\beta}) \dot{g}(\hat{\boldsymbol{\beta}}; x_2)}$$

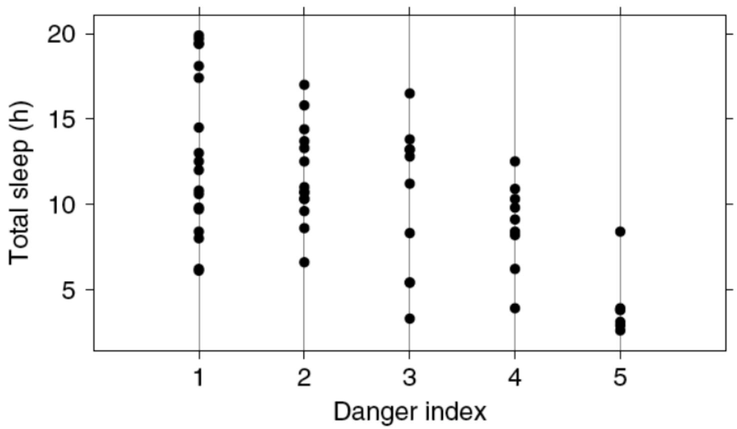
Factors

- allow qualitative or categorical predictors
- different levels: male or female, eye colour, etc.
- use dummy variables in the regression model
  not just one variable
- e.g., 0 for male and 1 for female, or $-1, 1$
- will give the same outcomes if you know what you are doing

• instead of the original categorical predictor

Factors - Sleep Data

How much time of sleeping/day



More danger less sleep

- sleep data - sleeping patterns of 62 mammal species (4 missing at random, thus omitted)
- response TS: total hours of sleep per day
- predictor D: danger indicator, 1 to 5, D=1 means least danger from other animals

in order

The Factor Rule

- the factor rule:

  A factor with d levels can be represented by at most d dummy variables. If the intercept is in the mean function, at most $d-1$ of the dummy variables can be used in the mean function

- define the j^{th} dummy variable $U_j, j = 1, \cdots, 5$

  only 1 at each level

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

overall 5 dummy variables get only 1 each row

- the regression model is:

$$\mathrm{E}(TS|D) = \beta_1 U_1 + \beta_2 U_2 + \beta_3 U_3 + \beta_4 U_4 + \beta_5 U_5 \rightarrow \text{rank } 5$$

5 columns
no intercept

Matrix with 0 1,
Not 12345

58 by 5
if use an intercept, by 5
not by 4

Two Models for the Same Thing

- β_j : can be interpreted as the population mean for all species with danger index j
- note that no intercept is there, why?
- now consider an equivalent model:

1 or 0

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

- $\eta_0 = \beta_1, \eta_0 + \eta_2 = \beta_2, \eta_0 + \eta_3 = \beta_3, \cdots, \eta_0 + \eta_5 = \beta_5$ number of hrs sleeping at different levels
- this is called a one-way analysis of variance model — fits a separate mean for each level

Model 6.1a

- (Table 6.1a) coefficient for U_j is the estimated mean for level j of D There are two tables

Stupid ANOVA

	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

no intercept

	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		

Model 6.1b

- (Table 6.1b) intercept: estimated mean for level 1 of D coefficient for U_j is the estimated difference between means for level 1 and level $j, j > 1$

	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

level 1
sleep more
so negative

	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		

< 0.3252
or 0.0675
or 0.0076

More on Models 6.1a and 6.1b

- how about the t-values?
- ANOVA Table 6.1a:

  NH: all β's are zero or $\mathrm{E}(TS|D) = 0$

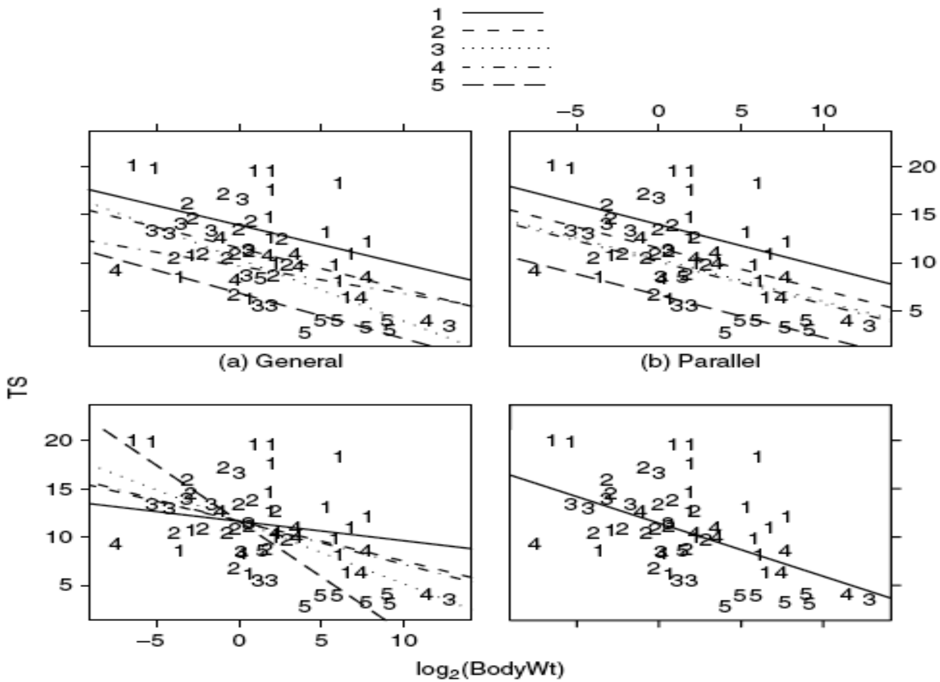
- ANOVA Table 6.1b:

  NH: $\mathrm{E}(TS|D) = \eta_0$

- caution: identical RSS's, the ANOVA in Table 6.1a is not an exclusive decomposition, $SYY \neq SS_{reg} + RSS$ without intercept. with intercept: same as SLR
- the 1st is easier to interpret, the 2nd is more used
- let's add a continuous predictor, $\log(\mathrm{BodyWt})$?

Adding a Continuous Term $\log(\text{BodyWt})$

- so two terms: D and $\log(\text{BodyWt})$
- four different cases



(a) General
(b) Parallel
(c) Common intercept
(d) Common regression

Model 1

- one regression line for each level of D
- $\mathrm{E}[TS|\log(\mathrm{BodyWt}), D] = \sum_{j=1}^{5}(\beta_{0j}U_j + \beta_{1j}U_j x)$
- $\mathrm{E}[TS|\log(\mathrm{BodyWt}), D] = \eta_0 + \eta_1 x + \sum_{j=2}^{5}(\eta_{0j}U_j + \eta_{1j}U_j x)$
- interactions between U_j and $\log(\mathrm{BodyWt})$
- first one is more convenient for obtaining interpretable parameters
- second one is useful for comparing mean functions
- what is the difference between this and fitting 5 separate regressions?

Other Models

- Model 2: parallel regression
- same slope but different intercepts, no interaction between U_j and $\log(\text{BodyWt})$
- when do we want to fit a model like this?
- Model 3: common intercept
- Model 4: coincident regression lines (no D)
- general F test: Model 1 as the model in AH
- NH: usually either Model 2 or 4
- what are the design matrices $\mathbf{X}$ for the above models?

Table 6.2

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

- exercise: compute F values from df and RSS
- more: ordinal factors sometimes may be treated as continuous, how to decide?