# STAT 3008 Applied Regression Analysis
# Tutorial 8.

XU Yongze & DONG Fangyuan

Nov 11 & 12 & 14, 2013

## Contents

# 1 Polynomial Regression

## 1.1 with one predictor

Model

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

- If a mean function with one predictor is smooth but not straight, integer powers of the predictors can be used to approximate it.
- e.g. quadratic regression, cubic regression.
- Any smooth function can be estimated by a polynomial of high-enough degree. Polynomials are generally used as approximations.

## 1.2 with several predictors

Model: e.g. second order mean function
♣ with interaction $x_1 x_2$:

$$\mathrm{E}(Y|X=x) = \beta_0 + \beta_1 x + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2$$
$$\mathrm{Var}(Y|X=x) = \sigma^2$$

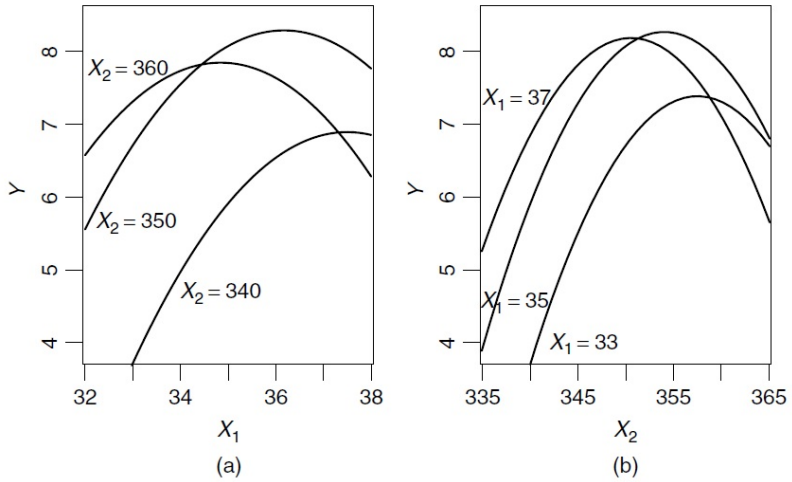


**FIG. 6.3** Estimated response curves for the cakes data, based on (6.7).

Each of the curves has a somewhat different shape. Effect of a change in $x_1$ depends on $x_2$.

♣ without interaction $x_1x_2$:

$$\mathrm{E}(Y|X=x) = \beta_0 + \beta_1 x + \beta_2 x_2 + \beta_{11}x_1^2 + \beta_{22}x_2^2$$
$$\mathrm{Var}(Y|X=x) = \sigma^2$$

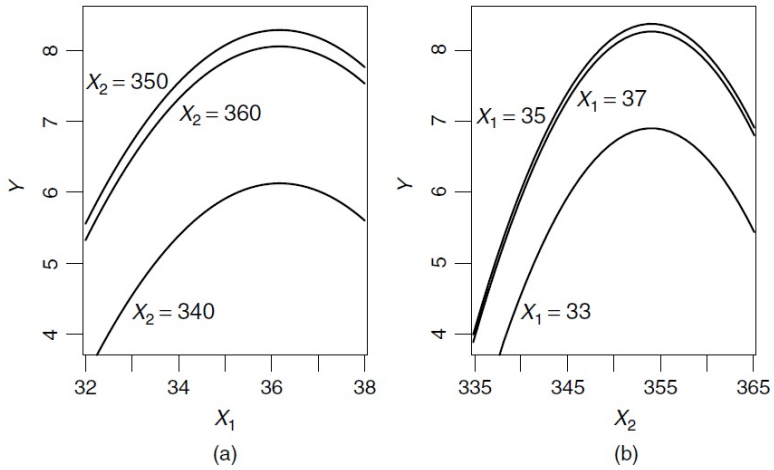


**FIG. 6.4** Estimated response curves for the cakes data, based on fitting with $\beta_{12} = 0$.

All curves within a plot have the same shape; all are maximized at the same point.

We can use the F-test to compare the models.

# 2 Delta method

Suppose we have:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

we can minimize or maximize $\hat{Y}$ with $\hat{X} = -\frac{\hat{\beta}_1}{2\hat{\beta}_2}$
We want to construct confidence interval for $\hat{X}$.

If we know the distribution of a r.v., we can find out 'lower quantile' and 'upper quantile' such that $P(\text{lower quantile} < r.v. < \text{upper quantile}) = 1 - \alpha$, so 'lower quantile ¡ r.v ¡ upper quantile' is the range we want.

We can apply the Delta Method.

♣ Known: $\hat{\theta} \sim N(\theta, \sigma^2 D)$
♣ Want: The distribution of $g(\hat{\theta})$

With Taylor expansion:

$$\begin{aligned} g(\hat{\theta}) &= g(\theta) + g'(\theta)^T(\hat{\theta} - \theta) + error \\ g(\hat{\theta}) &\approx g(\theta) + g'(\theta)^T(\hat{\theta} - \theta) \\ \mathrm{E}(g(\hat{\theta})) &\approx g(\theta) \\ \mathrm{Var}(g(\hat{\theta})) &\approx g'(\theta)^T \mathrm{Var}(\hat{\theta} - \theta) g'(\theta) = \sigma^2 g'(\theta)^T D g'(\theta) \end{aligned}$$

where

$$g'(\theta) = (\frac{\partial g}{\partial \theta_1}, \ldots, \frac{\partial g}{\partial \theta_k})^T$$

we have:

$$\boldsymbol{g(\hat{\theta}) \sim N(g(\theta), \sigma^2 g'(\theta)^T D g'(\theta))}$$

$$\frac{g(\hat{\theta} - g(\theta)}{\hat{\sigma}\sqrt{g'(\theta)^T D g'(\theta)}} \sim t(n - p - 1)$$

approximately,

$$\frac{g(\hat{\theta} - g(\theta)}{\hat{\sigma}\sqrt{g'(\theta)^T D g'(\theta)}} \sim N(0, 1)$$

then 95% C.I. for $g(\beta)$ is $g(\hat{\beta}) \pm 1.96\sqrt{\widehat{\mathrm{Var}(g(\hat{\beta}))}}$

# 3 Factor

## 3.1 Two type of variables

♣ Quantitative Variables

Measured in a numeric scale. The numbers used to value the variables are comparable. They can tell which one is 'larger' or 'smaller'. E.g. height, weight.

♣ Qualitative/Categorical Variables

Divided into levels or categories. E.g. gender, race, and blood type.

## 3.2 Dummy variable

♣ What is dummy variable?

Dummy variable is created so as to tackle regression analysis with qualitative variables. It can only take two values, usually, 0 and 1. It is usually an indicator variable I($Y_i$ in jth category).

♣ A factor with **d** level 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.

♣ For example,

Suppose there is a factor with 3 levels,

$$U_{ij} = I(Y_i \ in \ jth \ category) = \begin{cases} 1, & if \quad Y_i \quad in \quad jth \quad category \\ 0, & otherwise, \end{cases}$$

(Case with no intercept):
The model is:

$\mathrm{E}(Y|U) = \beta_1 U_1 + \beta_2 U_2 + \beta_3 U_3$

Suppose $U_1 = 1$, so $U_2 = U_3 = 0$. Therefore, $\mathrm{E}(Y|U_1 = 1) = \beta_1$.

Suppose $U_2 = 1$, so $U_1 = U_3 = 0$. Therefore, $\mathrm{E}(Y|U_2 = 1) = \beta_2$.

Suppose $U_3 = 1$, so $U_1 = U_2 = 0$. Therefore, $\mathrm{E}(Y|U_3 = 1) = \beta_3$.

$\beta_i$ can be interpreted as the population mean for all subjects with the ith category.

(Case with intercept):
The model is:

$\mathrm{E}(Y|U) = \alpha_0 + \alpha_2 U_2 + \alpha_3 U_3$

Suppose $U_1 = 1$, so $U_2 = U_3 = 0$. Therefore, $\mathrm{E}(Y|U_1 = 1) = \alpha_0$.

Suppose $U_2 = 1$, so $U_1 = U_2 = 0$. Therefore, $\mathrm{E}(Y|U_2 = 1) = \alpha_0 + \alpha_2$.

Suppose $U_3 = 1$, so $U_2 = U_3 = 0$. Therefore, $\mathrm{E}(Y|U_3 = 1) = \alpha_0 + \alpha_3$.

$\alpha_i$ can be interpreted as the difference between means for level 1 and level j for j ¿ 1.

Moreover, $\alpha_0 = \beta_1, \alpha_0 + \alpha_2 = \beta_2, \alpha_0 + \alpha_3 = \beta_3$

♣ Adding a continuous variable

general Regression:
$\mathrm{E}(Y|U, x) = \alpha_0 + \alpha_1 x + \sum_{j=2}^{d} (\alpha_{0j} U_j + \alpha_{1j} U_j x)$

Regression with common slope:(Parallel Regression)]
$\mathrm{E}(Y|U, x) = \alpha_0 + \alpha_1 x + \sum_{j=2}^{d} \alpha_{0j} U_j$

Regression with common intercept:
$\mathrm{E}(Y|U, x) = \alpha_0 + \alpha_1 x + \sum_{j=2}^{d} \alpha_{1j} U_j x$

Regression with common intercept and slope:(Coincident Regression)
$\mathrm{E}(Y|U, x) = \beta_0 + \beta_1 x$

# 4 Exercises

1. (from past Final) Under which characteristic(s) of residual plot, using WLS regression is useless?
   (i) The mean function is zero and the variance function is constant.
   (ii) The mean function is zero and the variance function is proportional to the value of reponse.
   (iii) The mean function is a sine function and the variance function is constant.
   (iv) The mean function is a sine function and the variance function is the value of response.

2. (from past Final) What is the name of the term $X_1X_2$ in the multiple linear regression model?

3. (from past Final) Consider the mean function: $E[Y\|U, X] = \beta_0 + \beta_1 U + \beta_2 X$
   When $U = \begin{cases} 0, & \text{if the observation is male} \\ 1, & \text{if the observation is female} \end{cases}$, then $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (1, 1, 1)$.
   If $U = \begin{cases} 2, & \text{if the observation is male} \\ 1, & \text{if the observation is female} \end{cases}$, then what is $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$?

4. (from past Final) A survey was conducted to study the relationship among the variables

   - Happiness(H) [1-10, higher score means happier]
   - Responsibility(R)
     - A: Extremely responsible personality
     - B: Moderately responsible personality
     - C: Faily responsible personality
     - D: Not reponsible at all
   - What's App usage(W)(messages/day)
   - Income(I)(HK$/month)
   - Gender(G)
     - M: Male
     - F: Female

   Seven people participated in the survey.

| Subject | Happiness(H) | Responsibility(R) | What's App(W) | Income(I) | Gender(G) |
|---|---|---|---|---|---|
| Issac | 9 | C | 0 | 13800 | M |
| Ralph | 1 | A | 90 | 16000 | M |
| Tom | 6 | B | 10 | 13800 | M |
| Fiona | 2 | A | 65 | 3000 | F |
| Kitty | 6 | B | 25 | 3500 | F |
| Bosco | 7 | A | 20 | 2000 | M |
| Rachael | 6 | B | 30 | 4000 | F |

(i) Which variabes are quantitative variables? Which variables are categorical variables?

(ii) Keith planned to do a regression between W and R. Following the textbook, he introduced three dummy variables for the four-level factor R and ran a regression of W against three dummy variables. However, he found something wrong in the R outputs.

- What was wrong? Explain this phenomeson and suggest modification methods.
- Outline the procedure to test if the What's App usages are different between an Extremely Responsible person and a Moderately responsible person.