Topic 8 - Multiple Regression

STAT 525 - Fall 2013

Outline

- Data and Notation
- Model (special cases)
- Estimation

The Data and Model

- Still have single response variable Y
- Now have multiple explanatory variables
- Examples:
  - Blood Pressure vs Age, Weight, Diet, Smoking, Fitness Level
  - Traffic Count vs Time, Location, Population, Month
- Goal: There is a total amount of variation in Y (SSTO). We want to explain as much of this variation as possible using a linear model and our explanatory variables

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- Have $p-1$ predictors $\longrightarrow$ p coefficients

Special Cases

- Polynomial of order $p-1$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_1 X_i^2 + \cdots + \beta_{p-1} X_i^{p-1} + \varepsilon_i$$

- Analysis of Variance
  - Predictors are sets of *indicator* or *dummy* variables (i.e., $X_{i,j}$=0 or 1)

- Interaction of predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

First Order Model with 2 Predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i; \quad i = 1, \ldots, n$$

- β_0 is the intercept and β_1 and β_2 are the regression coefficients
- Meaning of regression coefficients
  - β_1 describes change in <u>mean response</u> per unit increase in X_1 when X_2 is held constant
  - β_2 describes change in <u>mean response</u> per unit increase in X_2 when X_1 is held constant
- Variables X_1 and X_2 are additive. Value of X_1 does not affect the change due to X_2. There is no *interaction*.

Interaction Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

- Meaning of parameters:
  - Change in X_1 when $X_2 = x_2$

$$\begin{aligned} \Delta Y &= (\beta_0 + \beta_1(X_1+1) + \beta_2 x_2 + \beta_3(X_1+1)x_2) - \\ &\quad (\beta_0 + \beta_1 X_1 + \beta_2 x_2 + \beta_3 X_1 x_2) \\ &= \beta_1 + \beta_3 x_2 \end{aligned}$$

  - Change in X_2 when $X_1 = x_1$

$$\Delta Y = \beta_2 + \beta_3 x_1$$

- Rate of change for one variable affected by the other

Qualitative Predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

- Let $X_2 = 1$ if case from Purdue and $X_2 = 0$ otherwise
- Meaning of parameters:
  - Case from Purdue ($X_2 = 1$):

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 1 + \beta_3 X_1(1) \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \end{aligned}$$

  - Case from other location ($X_2 = 0$)

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 0 + \beta_3 X_1(0) \\ &= \beta_0 + \beta_1 X_1 \end{aligned}$$

- Have <u>two</u> regression lines
- β_2 and β_3 quantify the differences in intercepts and slopes

Response Surface

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

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

- Find $\mathbf{b}$ to minimize $(\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb})$
- $\mathbf{b} = (\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'Y}$
- Fitted values $\mathbf{HY}$ form response surface
- No longer a line but a (hyper)plane
- If $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I) \longrightarrow \mathbf{Y} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^{\mathbf{2}}\mathbf{I})$ which results in similar inference as before

First Order Model with 2 Predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i; \qquad i = 1, \ldots, n$$

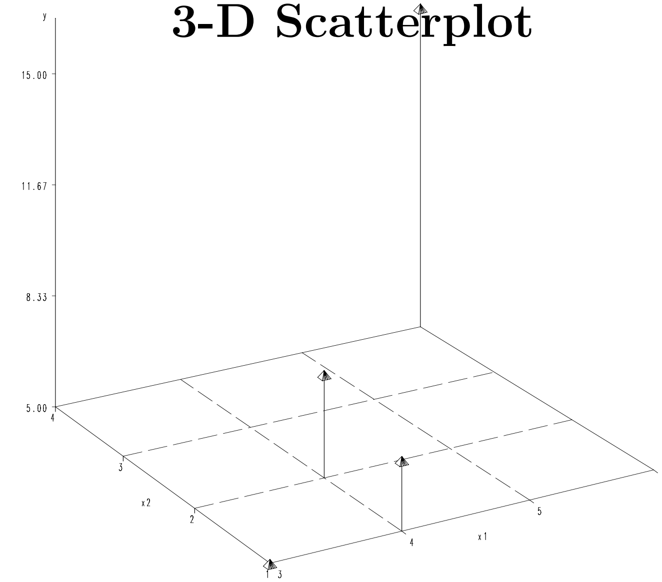
- Consider the following data set

Case	X_1	X_2	Y
1	3	1	5
2	4	2	8
3	4	1	7
4	6	4	15

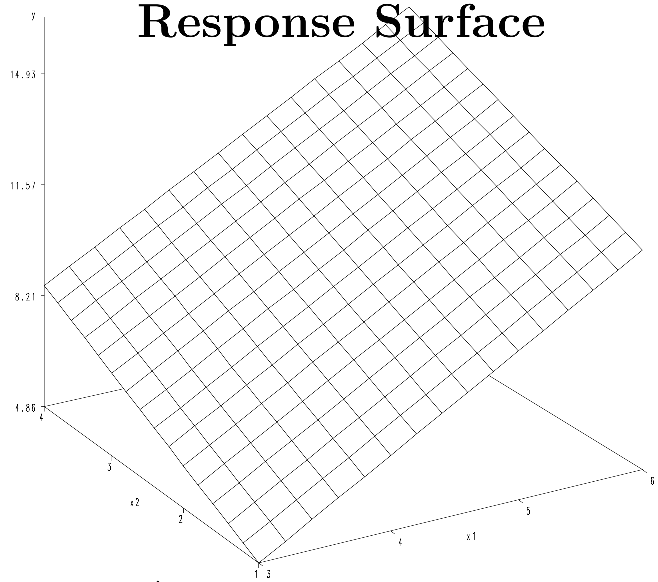
- Can show

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} -2.786 \\ 2.143 \\ 1.214 \end{bmatrix}$$

3-D Scatterplot

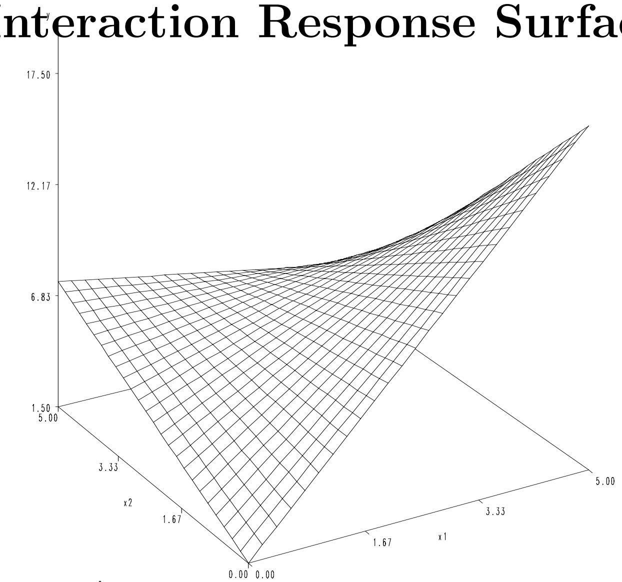


Response Surface



$\hat{Y}_i = -2.79 + 2.14X_{i1} + 1.21X_{i2}$

Interaction Response Surface



$\hat{Y}_i = 1.5 + 3.2X_{i1} + 1.2X_{i2} - .75X_{i1}X_{i2}$

Residuals

- $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$
- $\mathbf{I} - \mathbf{H}$ symmetric and idempotent
- Covariance Matrix

$$\begin{aligned} \boldsymbol{\sigma}^2(\mathbf{e}) &= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})' \\ &= \sigma^2(\mathbf{I} - \mathbf{H}) \end{aligned}$$

- $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$ where $h_{ii} = \mathbf{X_i'}(\mathbf{X'X})^{-1}\mathbf{X_i}$
- Residuals are usually correlated $(-\sigma^2 h_{ij})$

Estimation of σ^2

- Similar approach as before
- Now p model parameters

$$\begin{aligned} s^2 &= \frac{\mathbf{e'e}}{n-p} \\ &= \frac{(\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb})}{n-p} \\ &= \frac{\text{SSE}}{n-p} \\ &= \text{MSE} \end{aligned}$$

ANOVA TABLE

Source of Variation	df	SS	MS
Regression		SSR	SSR/
Error		SSE	SSE/
Total	$n-1$	SSTO	

- F Test: Tests whether there is a *regression* relation between the dependent variable Y and the *set* of predictors
  - $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$
  - H_a : at least one $\beta_k (k = 1, \ldots, p-1) \neq 0$
- Coefficient of Determination (R^2) describes proportionate reduction in variation of Y for the *set* of X variables

Background Reading

- KNNL Sections 6.1-6.5
- knnl216.sas
- KNNL Sections 6.6-6.7