Data modeling: CSCI E-106

Applied Linear Statistical Models

Chapter 6 – Multiple Regression I

Multiple Regression

- Multiple regression analysis is one of the most widely used of all statistical methods.
- a variety of multiple regression models
- basic statistical results in matrix form

First-Order Model with Two Predictor Variables

- Two predictor variables: X_1; X_2

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

- first-order model with two predictor variables: linear in the predictor variables.
  - Y_i denotes as usual the response in the ith trial,
  - X_{i1} and X_{i2} are the values of the two predictor variables in the ith trial.
  - Parameters of the model are β_o, β_1 and β_2, and the error term is ε_i.

- Assume $E\{\varepsilon_i\}= 0 \Rightarrow E\{Y\} = \beta_o + \beta_1 X_1 + \beta_2 X_2$

- $called\ a\ regression\ surface\ or\ a\ response\ surface$

First-Order Model with Two Predictor Variables, cont'd

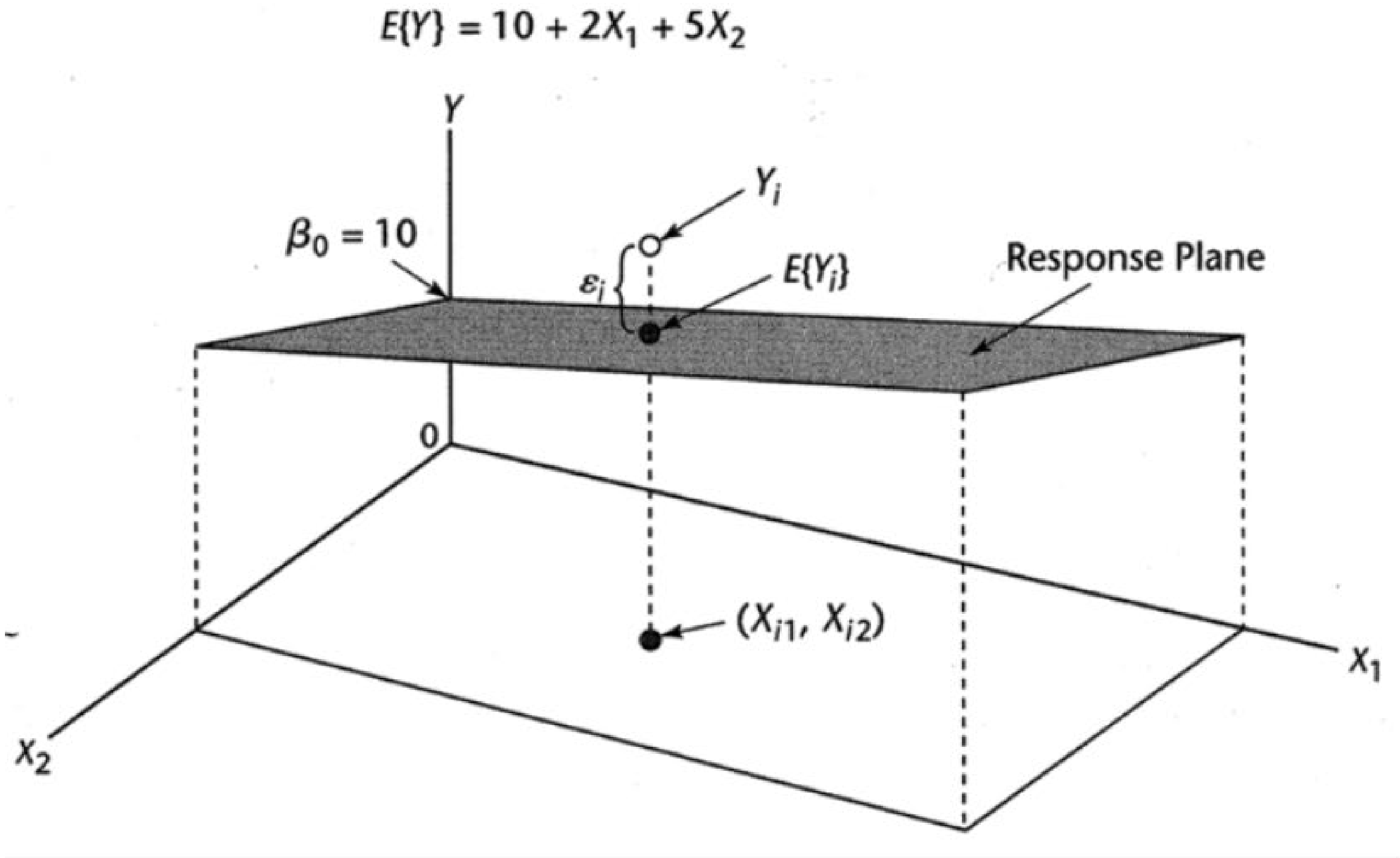


Figure : Response Function is a plane-Sales Promotion Example.

Meaning of Regression Coefficients

- The parameter β_1 indicates the change in the mean response $E\{Y\}$ per unit increase in X_1 when X_2 is held constant.
- Likewise, β_2 indicates the change in the mean response per unit increase in X_2 when X_1 is held constant
- When the effect of X_1 on the mean response does not depend on the level of X_2 , and correspondingly the effect of X_2 does not depend on the level of X_1, the two predictor variables are said to have additive effects or not to interact.
- The parameters β_1 and β_2 are sometimes called partial regression coefficients because they reflect the partial effect of one predictor variable when the other predictor variable is included in the model and is held constant.

Meaning of Regression Coefficients, cont'd

- β_0: intercept in the regression plane; $X_1 = 0, X_2 = 0$
- β_1: the change in the mean response with $\triangle X_1 = 1$ and X_2=constant; $\frac{\partial E\{Y\}}{\partial X_1} = \beta_1$
- β_2: the change in the mean response with X_1=constant and $\triangle X_2 = 1$; $\frac{\partial E\{Y\}}{\partial X_2} = \beta_2$
- $\beta_1,\ \beta_2$: partial regression coefficients $\because$ they reflect the partial effect
- $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$: additive or do not interact

First-Order Model with More than Two Predictor Variables

- The first-order regression model with *p -1* predictor variables:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \\
&= \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \varepsilon_i \\
&= \sum_{k=0}^{p-1} \beta_k X_{ik} + \varepsilon_i, \quad \text{where } X_{i0} \equiv 1
\end{aligned}
$$

- Assuming that $E\{\varepsilon_i\} = 0$, the response function (hyperplane) :

$$\Rightarrow E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}$$

First-Order Model with More than Two Predictor Variables, cont'd

- This response function is a hyperplane, which is a plane in more than two dimensions. It is no longer possible to picture this response surface
- β_k: the change in the mean response E{Y} with $\Delta X_k = 1$ and all other predictor variables are held constant
- additive and do not interact

General linear Regression Model

- The general linear regression model with normal error terms:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i,\ i = 1, \ldots, n \\
&= \beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad (X_{i0} \equiv 1) \\
&= \sum_{k=0}^{p-1} \beta_k X_{ik} + \varepsilon_i, \quad \text{where } X_{i0} \equiv 1
\end{aligned}
$$

- Parameters: β_0, β_1 ,…,β_{p-1}
- known constants: X_{i1},X_{i2},…,X_{ip-1}
- ε_i: independent $N(0,\sigma^2)$

General linear Regression Model, cont'd

- **p - 1 Predictor Variables**: When $X_1, \ldots, X_{p-1}$ represent p - 1 different predictor variables, there are no interaction effects between the predictor variables.
- **Qualitative Predictor Variables:** such as gender (male, female) or disability status(not disabled, partially disabled, fully disabled).
  - Indicator (dummy) variables are used to identify classes of a qualitative variable.
  - Example: Y is the length of hospital stay; X_1 is age, and X_2 is gender

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

X_{i1}: patient's age

$$X_{i2}: \begin{cases} 1, & \text{where patient female} \\ 0, & \text{where patient male} \end{cases}$$

Qualitative Predictor Variables

- The response function: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- The response function for male patients: $X_2 = 0$
  $E\{Y\} = \beta_0 + \beta_1 X_1$
- The response function for female patients: $X_2 = 1$

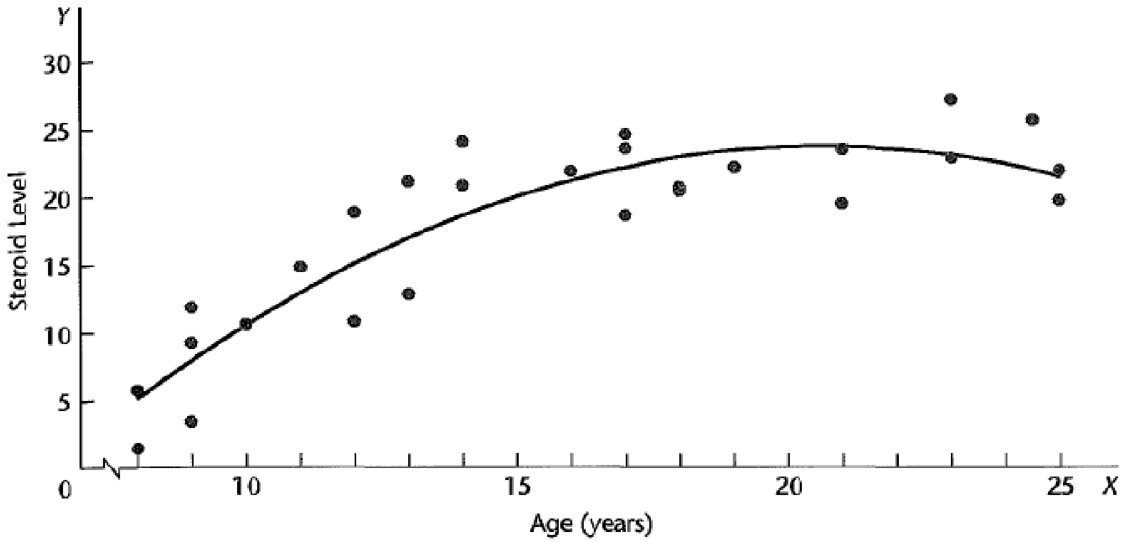
$$E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_1$$

Polynomial Regression

- Special cases of the general regression model
- Contain squared and higher order terms of the predicted variables makes the response function curvilinear.

$$Y_i = \beta_o + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

let $X_i = X_{i1}\ and\ X_i^2 = X_{i2} \Rightarrow Y_i = \beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$



Transformed Variables

- complex, curvilinear response functions, yet still are special cases of the general linear regression model.
- Examples:

$$\log Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \quad (Y_i' = \log Y_i)$$

$$Y_i = \frac{1}{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i} \quad (Y_i' = \tfrac{1}{Y_i})$$

Interaction Effects and Combination of Cases

An example of a nonadditive regression model with X_1, X_2: (complex)

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i \\
&= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i, \quad (X_{i3} = X_{i1} X_{i2})
\end{aligned}
$$

By cross-product term:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2 + \beta_3 X_{i2} + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \varepsilon_i \\
&= \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3} + \beta_4 Z_{i4} + \beta_5 Z_{i5} + \varepsilon_i
\end{aligned}
$$

Interaction Effects and Combination of Cases, cont'd

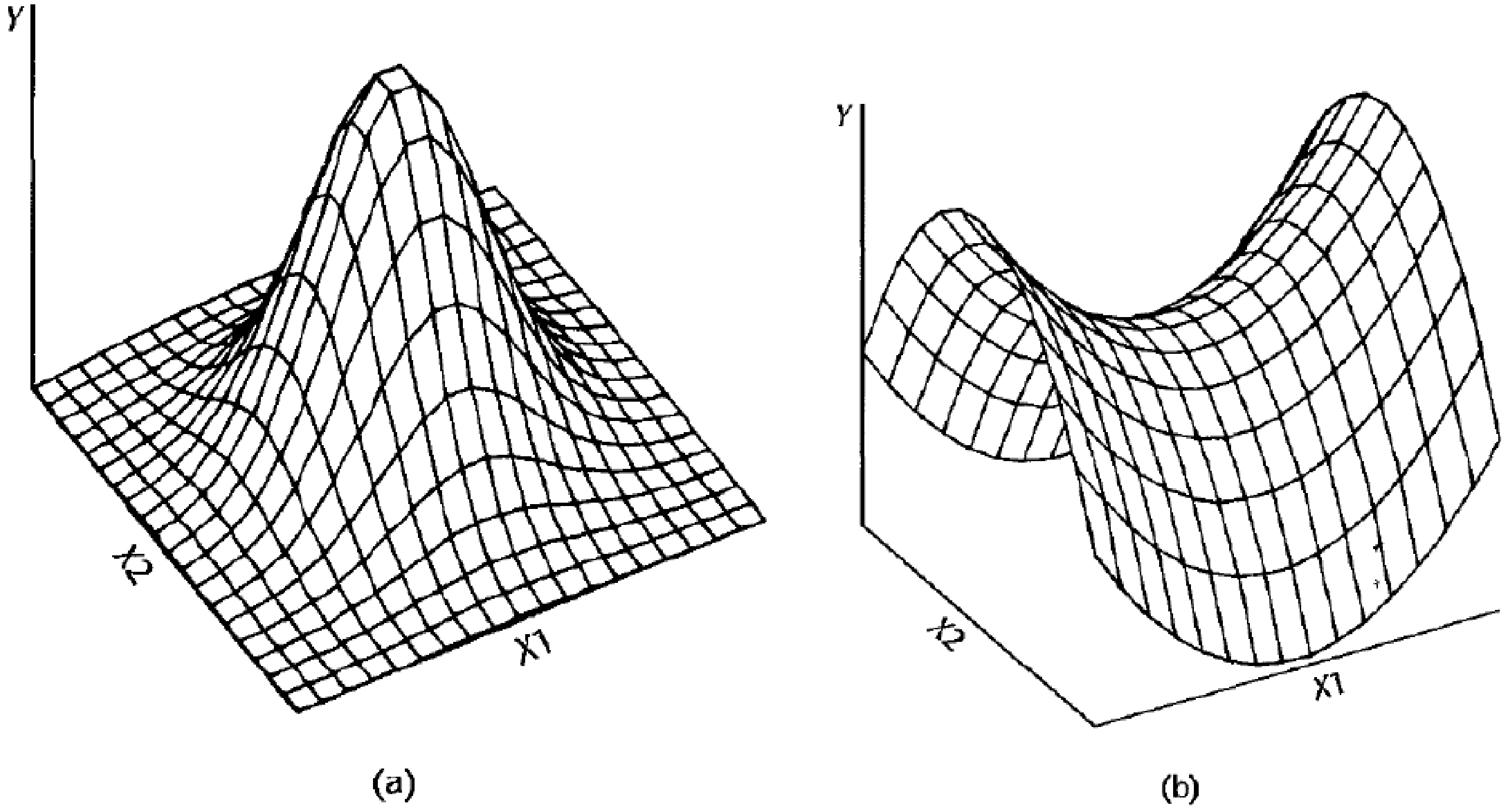


Figure : Additional Examples of Response Functions.

Meaning of Linear in General Linear Regression Model

- A regression model is linear in the parameters:

$$Y_i = c_{i0}\beta_0 + c_{i1}\beta_1 + c_{i2}\beta_2 + \cdots + c_{i,p-1}\beta_{p-1} + \varepsilon_i$$

where c_{ik}, $k = 0, \ldots, p-1$ are coefficients involving the predictor variables

- Illustration: nonlinear

$$Y_i = \beta_0 \exp(\beta_1 X_i) + \varepsilon_i$$

General Linear Regression Model in Matrix Terms

The general linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i,\ i = 1, \ldots, n$$

$$\underset{n\times 1}{\boldsymbol{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \underset{n\times p}{\boldsymbol{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

General Linear Regression Model in Matrix Terms, cont'd

$$\Rightarrow \underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times p}{\boldsymbol{X}}\,\underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

- $\boldsymbol{Y}$: vector of responses;
- $\boldsymbol{\beta}$: vector of parameters
- $\boldsymbol{X}$: matrix of constants
- $\boldsymbol{\varepsilon}$: vector of independent normal random variables;
  $\boldsymbol{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0}$; $\underset{n\times n}{\boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\}} = \sigma^2\boldsymbol{I}$
- Expectation and variance-covariance matrix of $\boldsymbol{Y}$:

$$\underset{n\times 1}{E\{Y\}} = X\beta; \quad \underset{n\times 1}{\sigma^2\{Y\}} = \sigma^2 I$$

Estimation of Regression Coefficients

- The general linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i,\ i = 1, \ldots, n$$

- The least squares criterion: the values $\beta_0, \ldots, \beta_{p-1}$ minimize Q

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1})^2$$

Estimation of Regression Coefficients, cont'd

- the vector of the least squares estimated regression coefficients:
- The normal equations:

$$\underset{p\times 1}{\boldsymbol{b}} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y}$$

- LSE:

$$\underset{p\times 1}{\boldsymbol{b}} = \underset{p\times p}{(\boldsymbol{X}'\boldsymbol{X})^{-1}} \underset{p\times 1}{\boldsymbol{X}'\boldsymbol{Y}}$$

Estimation of Regression Coefficients, cont'd

- The method of MLE leads to the same estimators of normal error regression model
- The likelihood function:

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1}X_{i,p-1})^2\right]$$

Maximizing the likelihood function with respect to $\beta_0, \beta_1, \ldots, \beta_{p-1}$ leads to the estimators $\boldsymbol{b}$

Fitted Values and Residuals

- The vector of $\hat{Y}_i$ and the vector of $e_i = Y_i - \hat{Y}_i$:

$$\underset{n\times 1}{\hat{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \boldsymbol{Xb} = \boldsymbol{HY} \quad (\underset{n\times n}{\boldsymbol{H}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')$$

$$\underset{n\times 1}{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{Xb} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \left(= (\boldsymbol{I} - \boldsymbol{H})\varepsilon \right)$$

Fitted Values and Residuals, cont'd

the variance-covariance matrix of the residuals:

$$\underset{n\times n}{\sigma^2\{\boldsymbol{e}\}} = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

$$\Rightarrow \text{(estimated by)} \quad \underset{n\times n}{s^2\{\boldsymbol{e}\}} = MSE(\boldsymbol{I} - \boldsymbol{H})$$

Analysis of Variance Results

- The sums of squares for ANOVA in matrix terms:

$$SSTO = \mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y} = \mathbf{Y}'\left[\ \mathbf{I} - \left(\tfrac{1}{n}\right)\mathbf{J}\ \right]\mathbf{Y}$$

$$SSE = \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b}) = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$SSR = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y} = \mathbf{Y}'\left[\ \mathbf{H} - \left(\tfrac{1}{n}\right)\mathbf{J}\ \right]\mathbf{Y}$$

$$MSR = \frac{SSR}{p - 1}$$

$$MSE = \frac{SSE}{n - p}$$

Analysis of Variance Results, cont'd

Table : ANOVA Table for General Linear Regression Model Model(6.19).

Source of Variation	SS	df	MS
Regression	$SSR = \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Y} - (\frac{1}{n})\boldsymbol{Y}'\boldsymbol{J}\boldsymbol{Y}$	$p-1$	$MSR = \frac{SSR}{p-1}$
Error	$SSE = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Y}$	$n-p$	$MSE = \frac{SSE}{n-p}$
Total	$SSTO = \boldsymbol{Y}'\boldsymbol{Y} - (\frac{1}{n})\boldsymbol{Y}'\boldsymbol{J}\boldsymbol{Y}$	$n-1$	

- $E\{MSE\} = \sigma^2$
- $p - 1 = 2$:

$$E\{MSR\} = \sigma^2 + \frac{1}{2}\Big[\beta_1^2 \sum (X_{i1} - \bar{X}_1)^2 + \beta_2^2 \sum (X_{i2} - \bar{X}_2)^2 + 2\beta_1\beta_2 \sum (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)\Big]$$

if $\beta_1 = 0 = \beta_2 \Rightarrow E\{MSR\} = \sigma^2$, otherwise $E\{MSR\} > \sigma^2$

F Test for Regression Relation

- Test whether there is a regression relation between Y and $\boldsymbol{X}$:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

$$H_a : \text{not all } \beta_k \ (k = 1, \ldots, p-1) \text{ equal zero}$$

$$\Rightarrow \ \text{test statistic: } F^* = \frac{MSR}{MSE}$$

- The decision rule to control the Type I error at α:

$$\text{If } F^* \leq F(1-\alpha; p-1, n-p), \ \text{conclude } H_0$$

$$\text{If } F^* > F(1-\alpha; p-1, n-p), \ \text{conclude} H_a$$

Coefficient of Multiple Determination

- The coefficient of multiple determination: R^2

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- Measures the proportionate reduction of total variation in Y associated with $X_1, \ldots, X_{p-1}$
- $0 \leq R^2 \leq 1$
- Adding more X variable$\Rightarrow R^2 \uparrow$
  - SSE can never become larger with more X variables
  - $SSTO$ is always the same of a given set of responses

Coefficient of Multiple Determination, cont'd

- The adjusted coefficient of multiple determination: R_a^2

$$R_a^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO}$$

  - R_a^2 becomes smaller when another X is introduce into the model ($\because$ $SSE \downarrow$)

- Coefficient of Multiple Correlation: R

$$R = \sqrt{R^2}$$

Inferences about Regression Parameters

Unbiased: $\boldsymbol{E}\{\boldsymbol{b}\} = \boldsymbol{\beta}$

The variance-covariance matrix $\sigma^2\{\boldsymbol{b}\}$:

$$\underset{p\times p}{\sigma^2\{\boldsymbol{b}\}} = \begin{bmatrix} \sigma^2\{b_0\} & \sigma\{b_0, b_1\} & \cdots & \sigma\{b_0, b_{p-1}\} \\ \sigma\{b_1, b_0\} & \sigma^2\{b_1\} & \cdots & \sigma\{b_1, b_{p-1}\} \\ \vdots & \vdots & & \vdots \\ \sigma\{b_{p-1}, b_0\} & \sigma\{b_{p-1}, b_1\} & \cdots & \sigma^2\{b_{p-1}\} \end{bmatrix}$$

$$= \underset{p\times p}{\sigma^2\{\boldsymbol{b}\}} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Inferences about Regression Parameters, cont'd

The estimated variance-covariance matrix $\mathbf{s}^2\{\boldsymbol{b}\}$:

$$\underset{p\times p}{s^2\{\boldsymbol{b}\}} = \begin{bmatrix} s^2\{b_0\} & s\{b_0, b_1\} & \cdots & s\{b_0, b_{p-1}\} \\ s\{b_1, b_0\} & s^2\{b_1\} & \cdots & s\{b_1, b_{p-1}\} \\ \vdots & \vdots & & \vdots \\ s\{b_{p-1}, b_0\} & s\{b_{p-1}, b_1\} & \cdots & s^2\{b_{p-1}\} \end{bmatrix}$$

$$= MSE(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Interval Estimation of β_k

- For the normal error regression model:

$$\frac{b_k - \beta_k}{s\{b_k\}} \sim t(n-p) \qquad k = 0, 1, \ldots, p-1$$

- The confidence limits for β_k with $1-\alpha$ confidence coefficient:

$$b_k \pm t(1-\alpha/2; n-p)s\{b_K\}$$

Interval Estimation of β_k ,cont'd

- Tests for β_k:

$$H_0 : \beta_k = 0 \quad H_a : \beta_k \neq 0$$

$$\Rightarrow t^* = \frac{b_k}{s\{b_k\}}$$

$\Rightarrow$ The decision rule:

If $|t^*| \leq t(1 - \alpha/2; n - p)$, conclude H_0

Otherwise conclude H_a

Joint Inferences

- The Bonferroni joint confidence intervals for g parameters with $1-\alpha$: ($g \leq p$)

$$b_k \pm Bs\{b_k\}$$
$$B = t(1-\alpha/2g; n-p)$$

(Chap. 7: tests concerning subsets of the regression parameters)

Interval Estimation of $E\{Y_h\}$

- Given values of $X_1, \ldots, X_{p-1}$: $X_{h,1}, \ldots, X_{h,p-1}$

$$\underset{p\times 1}{\boldsymbol{X}_h} = \begin{bmatrix} 1 \\ X_{h1} \\ \vdots \\ X_{h,p-1} \end{bmatrix}$$

- The mean response $E\{Y_h\}$:

$$E\{Y_h\} = \boldsymbol{X}_h'\boldsymbol{\beta}$$

Interval Estimation of $E\{Y_h\}$, cont'd

- The estimated mean response $\hat{Y}_h$:

$$
\begin{aligned}
\hat{Y}_h &= \boldsymbol{X}_h'\boldsymbol{b} \\
\Rightarrow E\{\hat{Y}_h\} &= \boldsymbol{X}_h'\boldsymbol{\beta} = E\{Y_h\} \quad (\textit{Unbiased}) \\
\sigma^2\{\hat{Y}_h\} &= \sigma^2\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h = \sigma^2\{\hat{Y}_h\} = \boldsymbol{X}_h'\boldsymbol{\sigma}^2\{\boldsymbol{b}\}\boldsymbol{X}_h
\end{aligned}
$$

- The estimated variance $s^2\{\hat{Y}_h\}$:

$$
s^2\{\hat{Y}_h\} = MSE(\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h) = \boldsymbol{X}_h'\boldsymbol{s}^2\{\boldsymbol{b}\}\boldsymbol{X}_h
$$

- The $1-\alpha$ confidence limits for $E\{Y_h\}$:

$$
\hat{Y}_h \pm t(1-\alpha/2; n-p)s\{\hat{Y}_h\}
$$

Confidence region for regression surface

- The Working-Hotelling confidence band for the regression line
- Boundary points of the confidence region at X_h:

$$\hat{Y}_h \pm W s\{\hat{Y}_h\}$$

$$W^2 = pF(1-\alpha; p, n-p)$$

Simultaneous Confidence Intervals for Several Mean Responses

- Estimate of $E\{Y_h\}$ corresponding to different X_h vectors with $1 - \alpha$:
  - Working-Hotelling confidence region bounds:

$$\hat{Y}_h \pm W s\{\hat{Y}_h\}$$

  - Bonferroni simultaneous confidence intervals: (g interval estimates)

$$\hat{Y}_h \pm B s\{\hat{Y}_h\}$$

$$B = t(1 - \frac{\alpha}{2g}; n - p)$$

Prediction of New Observation $Y_{h(new)}$

- The $1 - \alpha$ prediction limits for a new observation $Y_{h(new)}$ at X_h:
  - $\hat{Y}_h \pm t(1 - \frac{\alpha}{2}; n - p)s\{pred\}$
  - where $s^2\{pred\}$= MSE + $s^2\{\hat{Y}_h\}$= $MSE(1 + X_h'(X'X)^{-1}X_h)$

- Prediction of mean of m new observations at X_h:
  - $\hat{Y}_h \pm t(1 - \frac{\alpha}{2}; n - p)s\{predmean\}$
  - Where $s^2\{predmean\}$= $\frac{MSE}{m}$ + $s^2\{\hat{Y}_h\}$= $MSE(\frac{1}{m} + X_h'(X'X)^{-1}X_h)$

Prediction of g New Observation

- g new observations at g different levels X_h with family confidence coefficient $1 - \alpha$:

$$\hat{Y}_h \pm S s\{pred\}$$

where $S^2 = g\,F(1 - \alpha\,; g, n - p)$

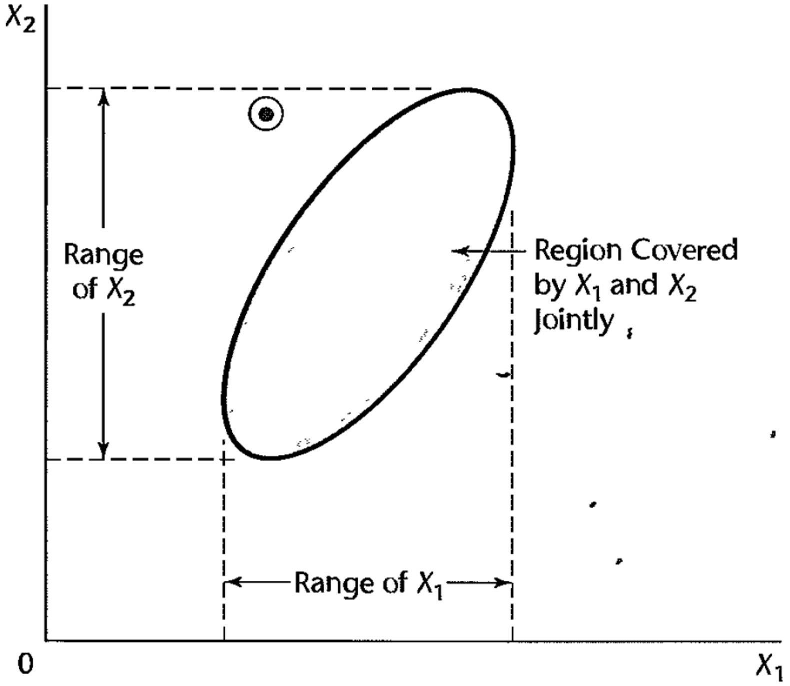
- Bonferroni simultaneous prediction limits:

$$\hat{Y}_h \pm B s\{pred\}$$

where $B = t\left(1 - \frac{\alpha}{2g}; n - p\right)$

Caution about Hidden Extrapolations

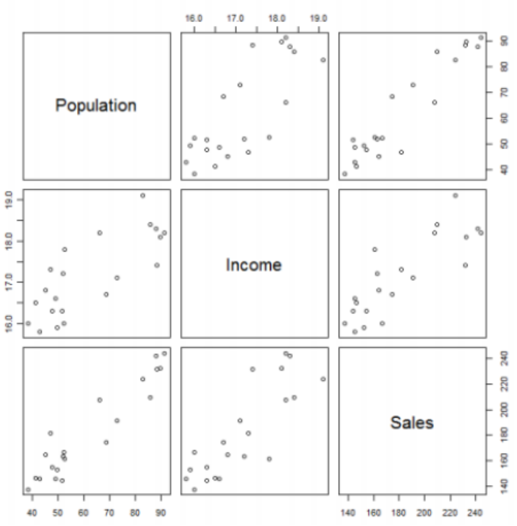
- Danger: the model may not be appropriate when it is extended outside the region of the observations.
- In multiple regression, it is particularly easy to lose track of this region since the levels of $X_1,\ldots, X_{p-1}$ jointly define the region. Thus, one cannot merely look at the ranges of each predictor variable.



  - The circled dot is within the ranges of the predictor variables X_1 and X_2 individually, yet is well outside the joint region of observations.
  - It is easy to spot this extrapolation when there are only two predictor variables, but it becomes much more difficult when the number of predictor variables is large.
  - We discuss in Chapter 10 a procedure for identifying hidden extrapolations when there are more than two predictor variables.

Diagnostics and Remedial Measures

- Diagnostics play an important role in the development and evaluation of multiple regression models.
- Most of the diagnostic procedures for simple linear regression that we described in Chapter 3 carry over directly to multiple regression.
- Some important ones will be discussed in Chapters 10 and 11.



(b) Correlation Matrix

	SALES	TARGTPOP	DISPOINC
SALES	1.000	.945	.836
TARGTPOP		1.000	.781
DISPOINC			1.000

Diagnostics and Remedial Measures, cont'd

- A complement to the scatter plot matrix that may be useful at times is the correlation matrix.
- This matrix contains the coefficients of simple correlation $r_{Y,X1}$, $r_{Y,X2}$, ... , $r_{Y,Xp-1}$ between Y and each of the predictor variables.
- As well as all of the coefficients of simple correlation among the predictor variables $r_{X1,X2}$, $r_{X1,X3}$, and etc.
- The format of the correlation matrix follows that of the scatter plot matrix:

$$\begin{bmatrix} 1 & r_{Y1} & r_{Y2} & \cdots & r_{Y,p-1} \\ r_{Y1} & 1 & r_{12} & \cdots & r_{1,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{Y,p-1} & r_{1,p-1} & r_{2,p-1} & \cdots & 1 \end{bmatrix}$$

Diagnostics and Remedial Measures, cont'd

- Scatter Plots
- Residual Plots
- Correlation Test for Normality
- Test for Constancy of Error Variance
- F Test for Lack of Fit
- Box-Cox tramsformations

Diagnostics and Remedial Measures, cont’d

All measures below discussed in Chapter 3 for simple linear regression can be carried over:

- Scatter Plots
- Residual Plots
- Correlation Test for Normality
- Test for Constancy of Error Variance
  - Brown-Forsythe Test
  - Breusch-Pagan Test
- F Test for Lack of Fit
- Box-Cox transformations

F test of lack of fit

As discussed Chapter 3, to test whether the multiple regression response function

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

is an appropriate response surface.

- Repeat observations
- SSE is decomposed into Pure Error (PE) and Lack of Fit (LOF) components

F test of lack of fit, cont'd

Testing:

$$H_0 : E\{Y\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$
$$H_a : E\{Y\} \neq \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

The test statistic:

$$F^* = \frac{SSLE}{c-p} \div \frac{SSPE}{n-c} = \frac{MSLF}{MSPE}$$
$$\Rightarrow \text{If } F^* \leq F(1-\alpha; c-p; n-c), \text{ conclude } H_0$$
$$\text{If } F^* > F(1-\alpha; c-p; n-c), \text{ conclude } H_a$$

Example: Dwaine Studio

```
> cor(DwaineStudio)
          X1        X2         Y
X1 1.0000000 0.7812993 0.9445543
X2 0.7812993 1.0000000 0.8358025
Y  0.9445543 0.8358025 1.0000000

> anova(f)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq  F value   Pr(>F)
X1         1 23371.8 23371.8 192.8962 4.64e-11 ***
X2         1   643.5   643.5   5.3108  0.03332 *
Residuals 18  2180.9   121.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

```
> f<-lm(Y~X1+X2,data=DwaineStudio)
> summary(f)

Call:
lm(formula = Y ~ X1 + X2, data = DwaineStudio)

Residuals:
     Min       1Q   Median       3Q      Max
-18.4239  -6.2161   0.7449   9.4356  20.2151

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.8571    60.0170  -1.147   0.2663
X1            1.4546     0.2118   6.868    2e-06 ***
X2            9.3655     4.0640   2.305   0.0333 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared:  0.9167,  Adjusted R-squared:  0.9075
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10
```

Example: Dwaine Studio, cont’d

- f1<-lm(Y~.,data=DwaineStudio) , will give you the previous slide model

- f1<-lm(Y~.^2,data=DwaineStudio), will give you the model with interaction term

- f1<-lm(Y~X1+X2+X1:X2,data=DwaineStudio), different way of fitting model with the interaction term