Multiple Regression

Yang Feng

Multiple regression

- One of the most widely used tools in statistical analysis
- Matrix expressions for multiple regression are the same as for simple linear regression

Need for Several Predictor Variables

Often the response is best understood as being a function of multiple input quantities

- Examples
  - Spam filtering-regress the probability of an email being a spam message against thousands of input variables
  - Revenue prediction - regress the revenue of a company against a lot of factors

First-Order with Two Predictor Variables

- When there are two predictor variables X_1 and X_2 the regression model

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

  is called a first-order model with two predictor variables.
- A first order model is linear in the predictor variables.
- X_{i1} and X_{i2} are the values of the two predictor variables in the i^{th} trial.

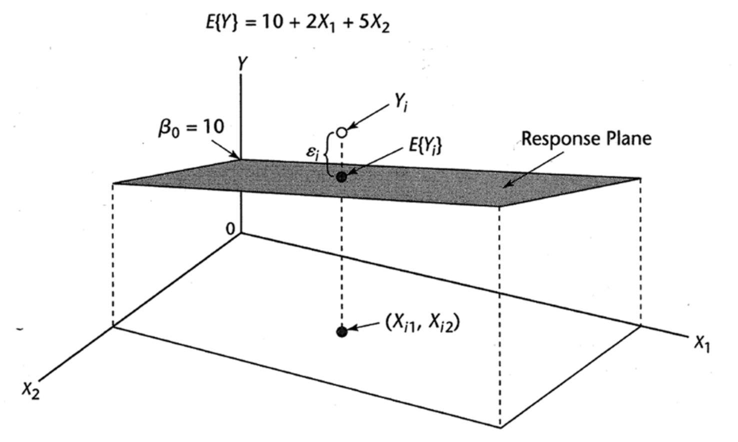
Functional Form of Regression Surface

- Assuming noise equal to zero in expectation

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- The form of this regression function is of a plane
  - e.g. $\mathbb{E}(Y) = 10 + 2X_1 + 5X_2$

Example



$E\{Y\} = 10 + 2X_1 + 5X_2$
Y
Y_i
$\beta_0 = 10$
ε_i
$E\{Y_i\}$
Response Plane
0
(X_{i1}, X_{i2})
X_1
X_2

Meaning of Regression Coefficients

- β_0 is the intercept when both X_1 and X_2 are zero;
- β_1 indicates the change in the mean response $\mathbb{E}(Y)$ per unit increase in X_1 when X_2 is held constant
- β_2 -vice versa
- Example: fix $X_2 = 2$

$$\mathbb{E}(Y) = 10 + 2X_1 + 5(2) = 20 + 2X_1 \qquad X_2 = 2$$

  intercept changes but clearly linear
- In other words, all one dimensional restrictions of the regression surface are lines.

Terminology

1. When the effect of X_1 on the mean response does not depend on the level X_2 (and vice versa) the two predictor variables are said to have *additive effects or not to interact*.
2. The parameters β_1 and β_2 are sometimes called *partial regression coefficients*. They represents the partial effect of one predictor variable when the other predictor variable is included in the model and is held constant.

Comments

1. A planar response surface may not always be appropriate, but even when not it is often a good approximate descriptor of the regression function in "local" regions of the input space
2. The meaning of the parameters can be determined by taking partials of the regression function w.r.t. to each.

First order model with > 2 predictor variables

Let there be $p-1$ predictor variables, then

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

which can also be written as

$$Y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \epsilon_i$$

and if $X_{i0} = 1$ is also can be written as

$$Y_i = \sum_{k=0}^{p-1} \beta_k X_{ik} + \epsilon_i$$

where $X_{i0} = 1$

Geometry of response surface

- In this setting the response surface is a *hyperplane*
- This is difficult to visualize but the same intuitions hold
  - Fixing all but one input variables, each β_p tells how much the response variable will grow or decrease according to that one input variable

General Linear Regression Model

We have arrived at the general regression model. In general the $X_1, ..., X_{p-1}$ variables in the regression model do not have to represent different predictor variables, nor do they have to all be quantitative(continuous).

The general model is

$$Y_i = \sum_{k=0}^{p-1} \beta_k X_{ik} + \epsilon_i \text{ where } X_{i0} = 1$$

with response function when $\mathbb{E}(\epsilon_i)$=0 is

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1}$$

Qualitative(Discrete) Predictor Variables

Until now we have (implicitly) focused on quantitative (continuous) predictor variables.

Qualitative(discrete) predictor variables often arise in the real world.

Examples:

- Patient sex: male/female
- College Degree: yes/no
- Etc

Example

Regression model to predict the length of hospital stay(Y) based on the age (X_1) and gender(X_2) of the patient. Define gender as:

$$X_2 = \begin{cases} 1 & \text{if patient female} \\ 0 & \text{if patient male} \end{cases}$$

And use the standard first-order regression model

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

Example cont.

- Where X_{i1} is patient's age, and X_{i2} is patient's gender
- If $X_2 = 0$, the response function is $E(Y) = \beta_0 + \beta_1 X_1$
- Otherwise, it's $E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1$
- Which is just another parallel linear response function with a different intercept

Polynomial Regression

- Polynomial regression models are special cases of the general regression model.
- They can contain squared and higher-order terms of the predictor variables
- The response function becomes curvilinear.
- For example $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$
  which clearly has the same form as the general regression model.

General Regression

- Transformed variables
  $\log Y, 1/Y$
- Interaction effects
  $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$
- Combinations
  $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \beta_4 X_{i1} X_{i2} + \epsilon_i$
- Key point-all linear in parameters!

General Regression Model in Matrix Terms

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ . \\ . \\ . \\ Y_n \end{pmatrix}_{n\times 1} \qquad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1,p-1} \\ \dots & & & & \\ 1 & X_{n1} & X_{n2} & \dots & X_{n,p-1} \end{pmatrix}_{n\times p}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ . \\ . \\ . \\ \beta_{p-1} \end{pmatrix}_{p\times 1} \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}_{n\times 1}$$

General Linear Regression in Matrix Terms

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

With $E(\boldsymbol{\epsilon}) = 0$ and

$$\sigma^2\{\boldsymbol{\epsilon}\} = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & & & \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

We have $E(\mathbf{Y}) = \mathbf{X}\beta$ and $\sigma^2\{\mathbf{Y}\} = \sigma^2\{\boldsymbol{\epsilon}\} = \sigma^2\mathbf{I}$

Least Square Solution

The matrix normal equations can be derived directly from the minimization of

$$Q(\mathbf{b}) = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b})$$

w.r.t. to $\mathbf{b}$

Key result

$$\frac{\partial \mathbf{X}\mathbf{b}}{\partial \mathbf{b}} = \mathbf{X}. \tag{1}$$

Least Square Solution

We can solve this equation

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

(if the inverse of $\mathbf{X}'\mathbf{X}$ exists) by the following

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

and since

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$$

we have

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Fitted Values and Residuals

Let the vector of the fitted values are

$$\hat{\mathbf{Y}} = \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ . \\ . \\ . \\ \hat{Y}_n \end{pmatrix}$$

in matrix notation we then have $\hat{\mathbf{Y}} = \mathbf{Xb}$

Hat Matrix-Puts hat on y

We can also directly express the fitted values in terms of $\mathbf{X}$ and $\mathbf{Y}$ matrices

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

and we can further define $\mathbf{H}$, the "hat matrix"

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} \qquad \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

The hat matrix plans an important role in diagnostics for regression analysis.

Hat Matrix Properties

1. the hat matrix is symmetric
2. the hat matrix is idempotent, i.e. $\mathbf{HH} = \mathbf{H}$

Important idempotent matrix property

For a symmetric and idempotent matrix $\mathbf{A}$, $rank(\mathbf{A}) = trace(\mathbf{A})$, the number of non-zero eigenvalues of $\mathbf{A}$.

Residuals

The residuals, like the fitted value $\hat{\mathbf{Y}}$ can be expressed as linear combinations of the response variable observations Y_i

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

also, remember

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{Xb}$$

these are equivalent.

Covariance of Residuals

Starting with

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

we see that

$$\sigma^2\{\mathbf{e}\} = (\mathbf{I} - \mathbf{H})\sigma^2\{\mathbf{Y}\}(\mathbf{I} - \mathbf{H})'$$

but

$$\sigma^2\{\mathbf{Y}\} = \sigma^2\{\epsilon\} = \sigma^2\mathbf{I}$$

which means that

$$\sigma^2\{\mathbf{e}\} = \sigma^2(\mathbf{I} - \mathbf{H})\mathbf{I}(\mathbf{I} - \mathbf{H}) = \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})$$

and since $\mathbf{I} - \mathbf{H}$ is idempotent (check) we have $\sigma^2\{\mathbf{e}\} = \sigma^2(\mathbf{I} - \mathbf{H})$

Quadratic Forms

- In general, a quadratic form is defined by

$$\mathbf{Y}'\mathbf{A}\mathbf{Y} = \sum_i \sum_j a_{ij} Y_i Y_j \text{ where } a_{ij} = a_{ji}$$

  with $\mathbf{A}$ the matrix of the quadratic form.

- The ANOVA sums SSTO,SSE and SSR can all be arranged into quadratic forms.

$$SSTO = \mathbf{Y}'(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{Y}$$

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

$$SSR = \mathbf{Y}'(\mathbf{H} - \frac{1}{n}\mathbf{J})\mathbf{Y}$$

Quadratic Forms

Cochran's Theorem

Let $X_1, X_2, \ldots, X_n$ be independent, $N(0, \sigma^2)$-distributed random variables, and suppose that

$$\sum_{i=1}^{n} X_i^2 = Q_1 + Q_2 + \ldots + Q_k,$$

where $Q_1, Q_2, \ldots, Q_k$ are nonnegative-definite quadratic forms in the random variables $X_1, X_2, \ldots, X_n$, with $rank(\mathbf{A}_i) = r_i$, $i = 1, 2, \ldots, k$. namely,

$$Q_i = \mathbf{X}'\mathbf{A}_i\mathbf{X}, \; i = 1, 2, \ldots, k.$$

If $r_1 + r_2 + \ldots + r_k = n$, then

1. $Q_1, Q_2, \ldots, Q_k$ are independent; and
2. $Q_i \sim \sigma^2\chi^2(r_i), \; i = 1, 2, \ldots, k$

ANOVA table for multiple linear regression

Source of Variation	SS	df	MS	$\mathbb{E}(MS)$
Regression	SSR	$p-1$	$MSR = SSR/(p-1)$	$> \sigma^2$
Error	SSE	$n-p$	$MSE = SSE/(n-p)$	σ^2
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$		

F-test for regression

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$
- H_a: no all $\beta_k, (k = 1, \cdots, p-1)$ equal zero

Test statistic:

$$F^* = \frac{MSR}{MSE}$$

Decision Rule:

- if $F^* \leq F(1-\alpha; p-1, n-p)$, conclude H_0
- if $F^* > F(1-\alpha; p-1, n-p)$, conclude H_a

R^2 and adjusted R^2

- The *coefficient of multiple determination* R^2 is defined as:

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

- $0 \le R^2 \le 1$
- R^2 always increases when there are more variables.
- Therefore, adjusted R^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 may decrease when p is large.
- *Coefficient of multiple correlation*:

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

Always positive square root!

Inferences about parameters

We have

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Since $\sigma^2\{\mathbf{Y}\} = \sigma^2\mathbf{I}$ we can write

$$\begin{aligned}
\sigma^2\{\mathbf{b}\} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{I} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

Also

$$\mathbb{E}(\mathbf{b}) = \mathbb{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbb{E}(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

Inferences

The estimated variance-covariance matrix

$$s^2\{\mathbf{b}\} = MSE(\mathbf{X}'\mathbf{X})^{-1}$$

Then, we have

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

- $1 - \alpha$ confidence intervals:

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

t test

- Tests for β_k:
  - $H_0 : \beta_k = 0$
  - $H_1 : \beta_k \neq 0$
- Test Statistic:

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

- Decision Rule:
  - $|t^*| \leq t(1 - \alpha/2; n - p)$; conclude H_0
  - Otherwise, conclude H_a

Joint Inferences

Bonferroni Joint Confidence Intervals for g parameters, the confidence limits with family confidence coefficient $1-\alpha$ are

$$b_k \pm Bs\{b_k\},$$

where

$$B = t(1-\alpha/(2g); n-p)$$

Interval estimate of EY_h

- We want to estimate the response at $\mathbf{X}_h = (1, X_{h1}, \cdots, X_{h,p-1})'$.
- Estimator: $\hat{Y}_h = \mathbf{X}_h'\mathbf{b}$
- Expectation $E\hat{Y}_h = \mathbf{X}_h'\beta = EY_h$
- Variance $\sigma^2\{\hat{Y}_h\} = \sigma^2\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h$
- Estimated Variance $s^2\{\hat{Y}_h\} = MSE(\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$
- $1-\alpha$ confidence limits for EY_h:

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

Confidence Region for Regression Surface and Prediction of New Observations

- Working-Hotelling confidence band:

$$\hat{Y}_h \pm Ws\{\hat{Y}_h\}$$

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

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

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

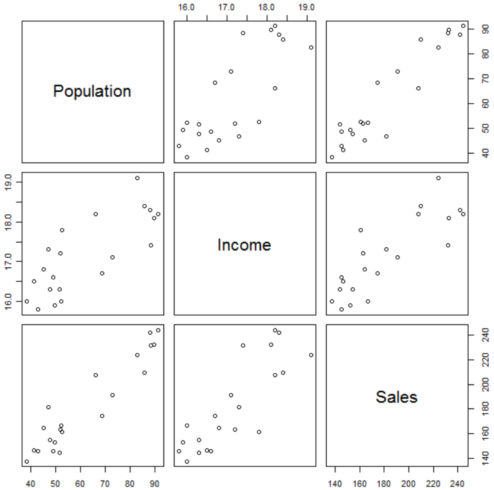
where $s^2\{pred\} = MSE + s^2\{\hat{Y}_h^2\} = MSE(1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$.

Diagnostics and Remedial Measures

- Very similar to simple linear regression.
- Only mention the difference.

Scatter Plot Matrix

- It is a matrix of scatter plot! R code: pairs(data)



Correlation Matrix

- Corresponds to the scatter plot matrix
- R code: cor(data)

	Population	Income	Sales
Population	1.00	0.78	0.94
Income	0.78	1.00	0.84
Sales	0.94	0.84	1.00

Other Diagnostics and Remedial Measures (Read after class)

- Residual Plots.
  - Against time (or some other sequence) for error dependency.
  - Against each X variable for potential nonlinear relationship and nonconstancy of error variances.
  - Against omitted variables (including the interaction terms). More on interaction terms in next Chapter.
- Correlation Test for Normality (Same, since it is on the residuals)
- Brown-Forsythe Test for Constancy of Error Variance (Need to find a way to divide the $\mathbf{X}$ space)
- Breusch-Pagan Test for Constancy of Error Variance (Same)
- F Test for Lack of Fit (Need to have (near) replicates on all dimension of $\mathbf{X}$)
- Box-Cox Transformations (Same, since it is on Y)