STAT 408
Applied Regression Analysis

Miles Xi

Department of Mathematics and Statistics

Loyola University Chicago

Fall 2022

Multiple Linear Regression

Motivation

- In reality, it is very rare to include only one predictor in linear model
  - Smoking habit, income, heath condition may affect the baby's birth weight
  - The insulin level may relate to both glucose and test result (diabetes or not)

- If a linear model includes more than one predictor, we call it <u>multiple linear regression</u>

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

where $p > 2$

p: number of parameters

$p-1$: number of predictors

Multiple Linear Regression

- The estimation of parameters in multiple linear regression is still least square estimation
- Suppose the data is

observation 1: $(x_{11}, x_{12}, \dots, x_{1(p-1)}, y_1)$

observation 2: $(x_{21}, x_{22}, \dots, x_{2(p-1)}, y_2)$

…

observation n: $(x_{n1}, x_{n2}, \dots, x_{n(p-1)}, y_n)$

- The residual sum of square is

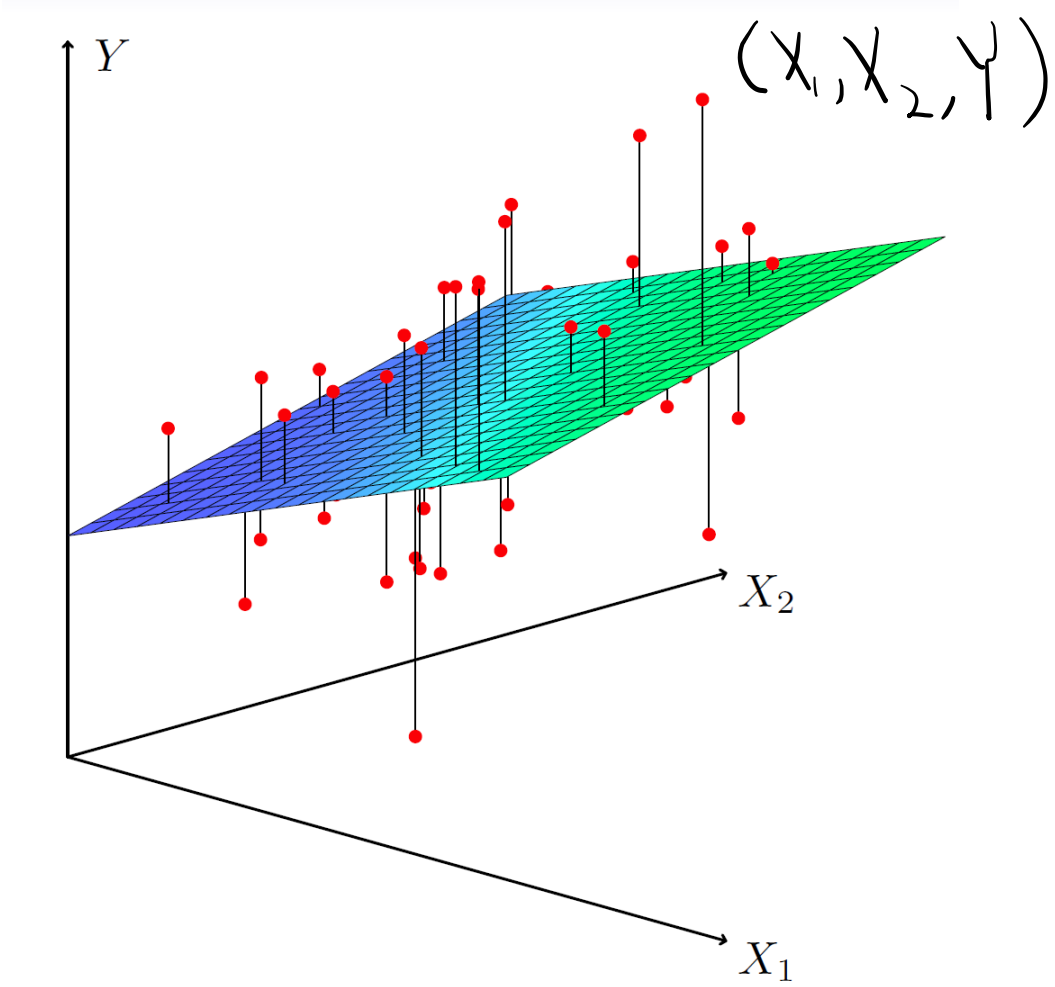
$$RSS = \sum_{i=1}^{n} \left(y_i - \underbrace{\beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_{p-1} x_{i(p-1)}}_{\hat{y}_i}\right)^2$$

Multiple Linear Regression

- We want to find parameters $\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2, \dots, \hat{\beta}_{p-1}$ such that RSS is minimized

$$\begin{cases} \dfrac{\partial RSS}{\partial \beta_0} = 0 \\ \dfrac{\partial RSS}{\partial \beta_1} = 0 \\ \dfrac{\partial RSS}{\partial \beta_2} = 0 \\ \dots \\ \dfrac{\partial RSS}{\partial \beta_{p-1}} = 0 \end{cases}$$

Multiple Linear Regression



Visualization of linear model with two predictors

Matrix Representation of Multiple Linear Regression

- However, the simple algebra notation doesn't work well in multiple linear model
- To fully explore the multiple linear model for high dimension data, we will use matrix representation

- Suppose we have a response variable Y and $p-1$ predictors $X_1, \dots, X_{p-1}$, where $p > 2$
- If the data size is n, then the data can be presented in the matrix form

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1(p-1)} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{n(p-1)} \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

where X is an $n \times (p-1)$ matrix, and y is an n dimensional vector

Matrix Representation of Multiple Linear Regression

- For each observation i $(x_{i1}, x_{i2} \ldots, x_{i(p-1)}, y_i)$, the multiple linear model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i(p-1)} + \epsilon_i$$

- Define the parameter vector and error vector

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Matrix Representation of Multiple Linear Regression

- Let's add one column to the matrix X

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1(p-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n(p-1)} \end{pmatrix}$$

Now the X is an $n \times p$ matrix (also called design matrix)

- With all the matrix notations, the matrix form of multiple linear model is

$$y = X\beta + \varepsilon$$

Least Square Estimation

- To estimate the model parameters, we still want to minimize the residual sum of squares, but in the matrix form

$$RSS(\beta) = \sum_{i=1}^{n} e_i^2 = e^T e$$

where T is matrix transpose, e is the residual vector

$$e = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} y_1 - \beta_0 - \beta_1 x_{11} - \cdots - \beta_p x_{1(p-1)} \\ \vdots \\ y_n - \beta_0 - \beta_1 x_{n1} - \cdots - \beta_p x_{n(p-1)} \end{pmatrix} = y - X\beta$$

Least Square Estimation

- Again, we need to take derivative of $RSS(\beta)$ in respect to β, and let it equal to zero

$$\frac{\partial e^T e}{\partial \beta} = \frac{\partial (y - X\beta)^T (y - X\beta)}{\partial \beta} = 0$$

- In this derivative, the numerator is a scaler and denominator is a $p \times 1$ vector

- The derivative result is following the same dimension as denominator $p \times 1$

Least Square Estimation

- Therefore, the optimization in matrix form

$$\frac{\partial e^T e}{\partial \beta} = \frac{\partial (y - X\beta)^T (y - X\beta)}{\partial \beta} = 0$$

is equivalent to the system of equations

$$\begin{cases} \frac{\partial RSS}{\partial \beta_0} = 0 \\ \frac{\partial RSS}{\partial \beta_1} = 0 \\ \cdots \\ \frac{\partial RSS}{\partial \beta_{p-1}} = 0 \end{cases}$$

Least Square Estimation

- Now the question is, how to take derivative in the matrix form?
  - We will not review the whole linear algebra class, but give some tips

- Recall in scaler case, if $f(x) = x^2$

$$\frac{df(x)}{dx} = 2x$$

- if $f(x) = cx$, where c is a constant

$$\frac{df(x)}{dx} = c$$

- The derivative in matrix form has similar rules

Least Square Estimation

Tip 1. If M is a matrix, then $M^T M$ is similar to M^2

Tip 2. If a is a constant vector or matrix, then Ma is similar to the previous scaler cx

With those two rules, the derivative is

$$\frac{\partial (y-X\beta)^T (y-X\beta)}{\partial \beta} \rightarrow \frac{\partial (y-X\beta)^2}{\partial \beta} = 2(y - X\beta) \times \frac{\partial (-X\beta)}{\partial \beta} = -2X^T (y - X\beta)$$

Tip 3. All matrix product after derivative must match dimension

Least Square Estimation

Now we have

$$\frac{\partial \varepsilon^T \varepsilon}{\partial \beta} = \frac{\partial (y - X\beta)^T (y - X\beta)}{\partial \beta} = -X^T 2(y - X\beta) = -2X^T y + 2X^T X\beta = 0$$

Cancel 2, we have

$$X^T X\beta = X^T y$$

Suppose $X^T X$ is invertible, multiply $(X^T X)^{-1}$ on both side gives

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Least Square Estimation

If M is a matrix, then its inverse matrix M^{-1} is defined as

$$MM^{-1} = M^{-1}M = I$$

where I is identity matrix

Tip 4. Identity matrix is the “scaler one” in matrix form; inverse is “scaler division” in matrix form

- $\hat{\beta} = (X^T X)^{-1} X^T y$ is the least square estimation for multiple linear model

Least Square Estimation

- The model fitted response is

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$$

where matrix H is called hat matrix (what's the dimension?)

- The residual is

$$e = y - \hat{y} = y - Hy = (I - H)y$$

Least Square Estimation

- Interesting property about hat matrix H

$$H^T = \left[X\left(X^TX\right)^{-1}X^T\right]^T = H$$

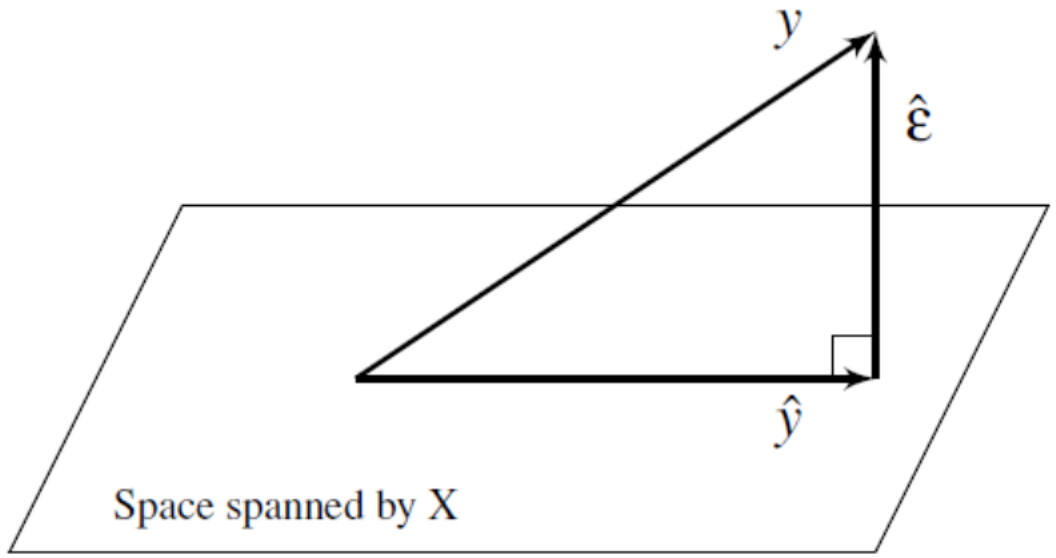
$$H^TH = X\left(X^TX\right)^{-1}X^T\ X\left(X^TX\right)^{-1}X^T = X\left(X^TX\right)^{-1}X^T = H$$

- Hat matrix is symmetric; the square of hat matrix is itself
- The RSS is

$$e^Te = [(I-H)y]^T[(I-H)y] = y^T(I-H)^T(I-H)y = y^T\left(I-H-H^T+H^TH\right)y$$

$$= y^T(I-H)y$$

Geometric interpretation



- Consider our linear model

$$y = X\beta + \varepsilon$$

- The response y is a vector in n dimensional space
- $X\beta$ is the space spanned by columns of X
- We want to find $\hat{\beta}$ such that the projection of y on X space is as close to y as possible
- The best choice is the $\hat{\beta}$ which makes $\hat{y}$ orthogonal to residual e
- Remember $\hat{y} = Hy$, H is also called orthogonal projection matrix

Geometric interpretation

- Essentially, the linear model is to represent complex n-dimensional y, using simpler p dimensional predictors (including intercept)

- The information in the data should be captured in those p dimensions

- Other information (random variation) is left in the $(n-p)$ dimensional residuals

$$\begin{aligned} \text{Data} &= \text{Systematic Structure} + \text{Random Variation} \\ n \text{ dimensions} &= p \text{ dimensions} + (n-p) \text{ dimensions} \end{aligned}$$

- $n-p$ is called the degree of freedom in linear model

One example

- In pima dataset, let's treat insulin as response variable Y, and other 8 variables as predictors

```
# read data
pima <- read.csv('pima.csv')

# remove missing values
pima <- pima[complete.cases(pima), ]

# multinormial linear regression
lm.model <- lm(insulin~., data = pima)
summary(lm.model)
```

One example

- Now let's fit a linear model by using $\hat{\beta} = (X^T X)^{-1} X^T y$

- First, manually construct X and response variable Y

```
X <- model.matrix(~pregnant+glucose+diastolic+triceps+bmi+diabetes+age+test, data = pima)
```

- Second, calculate $(X^T X)^{-1}$

```
XtXi <- solve(t(X)%*%X)
```

- Third, calculate $\hat{\beta} = (X^T X)^{-1} X^T y$

```
XtXi%*%t(X)%*%y
```

Goodness of Fit

- Recall that R^2 is defined as the fraction of variance explained by the model

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

where

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- What is the potential issue of using R^2 as goodness of fit in multiple linear regression?

Adjusted R^2

- Adding predictors will always “explain” some variations, even if the predictor is random and not related to response variable

- In multiple linear regression, we have to consider the number of predictors in the model (model complexity)

- Adjusted R^2 is the “average” fraction of variance explained per predictor

$$R_a^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}$$

- Essentially, we normalize the sum of square by corresponding number of parameters

Adjusted R^2

$$R_a^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}$$

- When adding one predictor X_j, both RSS and $n - p$ decrease
  1. If X_j is significantly related to Y, then RSS decreases "faster" than $n - p$, $RSS/(n-p) \downarrow$, $R_a^2 \uparrow$
  2. If X_j is not significantly related to Y, then RSS decreases "slower" than $n - p$, $RSS/(n-p) \uparrow$, $R_a^2 \downarrow$

Adjusted R^2

- We can show that

$$R_a^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)} = 1 - (\frac{n-1}{n-p})(1-R^2)$$

- Since $p > 2$ in multiple linear regression, $n - p < n - 1$, which means $R_a^2 < R^2$ (always, check R code output)

- Adjusted R^2 considers the model complexity and penalizes larger model

Identifiability

- The least square estimate $\hat{\beta} = (X^T X)^{-1} X^T y$ relies on the the successful inverse of $X^T X$

- If $X^T X$ is singular (not full rank), then there will be infinitely many solutions to the equation

$$X^T X\beta = X^T y$$

which is called unidentifiable

Identifiability

- To understand this, imagine in a system of equations, the number of unique functions is less than the number of variables

$$\begin{cases} x + y = 1 \\ 2x + 2y = 2 \end{cases}$$

$$\begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix} * \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

- Then the number of unique equations passes the number of unknown parameters
- It has infinite number of solutions (no unique solution)

Identifiability

- In multiple linear regression, unidentifiability occurs when

1. X's columns are linearly dependent
   - A person's weight is measured both in pounds and kilos ($X_1 = 0.45X_2$)
   - For each individual, we record the number of years of preuniversity education, the number of years of university education, and the total number of years of education ($X_1 = X_2 + X_3$)

2. There are more parameters than observations $p > n$

Identifiability

- R automatically fits the largest identifiable model by removing perfectly dependent variables

- Suppose we create a new variable for the pima dataset

```
pima$new <- pima$bmi+pima$age
lm.model <- lm(insulin~., data = pima)
summary(lm.model)
```

Identifiability

- More severe issue happens if we are close to unidentifiability
- Suppose we add a small random perturbation to variable glucose by adding a random variate from a uniform distribution U [−0.0005,0.0005]
- This will break the exactly linear relationship, but it is still close to perfect

Identifiability

```
# read data
pima <- read.csv('pima.csv')

# remove missing values
pima <- pima[complete.cases(pima), ]

# add small perturbation
pima$glucose.p <- pima$glucose + 0.001*(runif(dim(pima))-0.5)

# fit linear model
lm.model <- lm(insulin~., data = pima)
summary(lm.model)
```

Identifiability

- All parameters are estimated, but the standard errors are very large, compared to the original data
- We cannot estimate them in a stable way
- For any "new" data from the same population, the corresponding "new" $\hat{\beta}$ will be very different
- It is harder to reject H_0 in t test, even though H_0 is wrong, because the noise is larger