Linear Regression-4

Prof. Asim Tewari
IIT Bombay

Multiple Linear Regression

Multiple Linear Regression assumes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

The model can be expressed as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

with its coefficients being derived by minimizing RSS

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

$$\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Multiple Linear Regression

If X is a vector $\begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^p \end{bmatrix}$

Data : n-data points

$$(x_1^T, y_1), (x_2^T, y_2) \cdots\cdots (x_n^T, y_n)$$

$$\begin{bmatrix} x_1^1 \\ x_1^2 \\ x_1^3 \\ \vdots \\ x_1^p \end{bmatrix} \qquad \begin{bmatrix} x_n^1 \\ x_n^2 \\ \vdots \\ x_n^p \end{bmatrix}$$

Multiple Linear Regression

Data: $(x_1, y_1), \ldots (x_i, y_i) \ldots (x_n, y_n)$

$$\underset{p\times 1}{x_i} = (x_i^1, x_i^2 \ldots x_i^p)^T$$

$$\underset{(p+1)\times 1}{\beta} = (\beta_0, \beta_1, \beta_2 \ldots \beta_p)^T$$

$$\underset{(p+1)\times 1}{x_i} (1, x_i^1, x_i^2, x_i^3 \ldots x_i^p)$$

Multiple Linear Regression

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{n \times 1} = \underbrace{\underbrace{\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}}_{n \times (p+1)} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}}_{(p+1) \times 1}}_{n \times 1}$$

$$\underbrace{Y}_{n \times 1} = \underbrace{X}_{n \times (p+1)} \underbrace{\beta}_{(p+1) \times 1}$$

$$x_1^T \rightarrow \begin{bmatrix} 1 \\ x_1^1 \\ x_1^2 \\ \vdots \\ x_1^p \end{bmatrix}$$

$p \rightarrow$ # of features

$n =$ # of samples

Multiple Linear Regression

$$RSS = \| \underbrace{X\beta}_{\hat{Y}} - Y \|^2$$

$$RSS = f(\beta)$$

Find β^* such that $RSS(\beta^*)$ is minimum.

Solve $\nabla RSS(\beta^*) = 0$ to get β^*.

Multiple Linear Regression

$$RSS(\beta) = \|X\beta - Y\|^2 = (X\beta - Y)^T (X\beta - Y)$$

$$= (X\beta)^T X\beta - (X\beta)^T Y - Y^T X\beta + Y^T Y$$

$$= \beta^T X^T X \beta - 2\beta^T X^T Y + Y^T Y$$

$$\left[\nabla_x (a^T x) = a \text{ and } \nabla_x (x^T A x) = (A + A^T) x \right]$$

Multiple Linear Regression

$$RSS(\beta) = \beta^T X^T X\beta - 2\beta^T X^T Y + Y^T Y$$

$$\nabla_\beta RSS(\beta) = \left(X^T X + X^T X\right)\beta - 2X^T Y$$

$$\left[\begin{array}{l} \nabla_x (a^T x) = a \\ \nabla_x (x^T A x) = \\ (A + A^T) x \end{array}\right]$$

$$= 2X^T X\beta - 2X^T Y$$

$\therefore$ Solving $\nabla_\beta RSS(\beta^*) = 0$ for β^*

$$2X^T X\beta^* - 2X^T Y = 0$$

$$\Rightarrow \boxed{\underbrace{\beta^*}_{(p+1)\times 1} = (X^T X)^{-1}(X^T Y)}$$

Multiple Linear Regression

$$\beta^* = (X^T X)^{-1} X^T Y$$

↳ Is this min or max?

Is this local or global?

$$\therefore \quad \nabla_\beta RSS(\beta) = 2 X^T X \beta - 2 X^T y$$

$$\therefore \quad \nabla^2 RSS(\beta) = 2 \underbrace{\underbrace{X^T}_{(p+1)n} \underbrace{X}_{n \times (p+1)}}_{(p+1) \times (p+1)} - 0$$

Multiple Linear Regression

$$\underbrace{\nabla^2 RSS(\beta)}_{\text{Hessian of RSS}} = 2X^T X$$

$$\forall \beta \quad \beta^T (2X^T X)\beta = 2(X\beta)^T X\beta$$

$$= 2\|X\beta\|^2 \geq 0$$

$\therefore \quad \beta^*$ is a global min

Multiple Linear Regression

$$\beta^* = \underbrace{(X^T X)^{-1}}\ \underbrace{X^T Y}$$

$$X = \begin{bmatrix} 1 & x_1^1 & x_1^2 & \cdots & x_1^p \\ 1 & & & & \\ \vdots & & & & \\ 1 & x_n^1 & x_n^2 & \cdots & x_n^p \end{bmatrix}$$

$n \times (p+1)$

$$\underbrace{(p+1) \times n \quad n \times (p+1)}$$

$$(p+1)\ (p+1)$$

$$\underbrace{(p+1) \times n \quad n \times 1}$$

$$(p+1)\ 1$$

$$(p+1) \times 1$$

Linear Regression

$\nabla_\beta RSS(\beta) = 2X^T X\beta - 2X^T y$

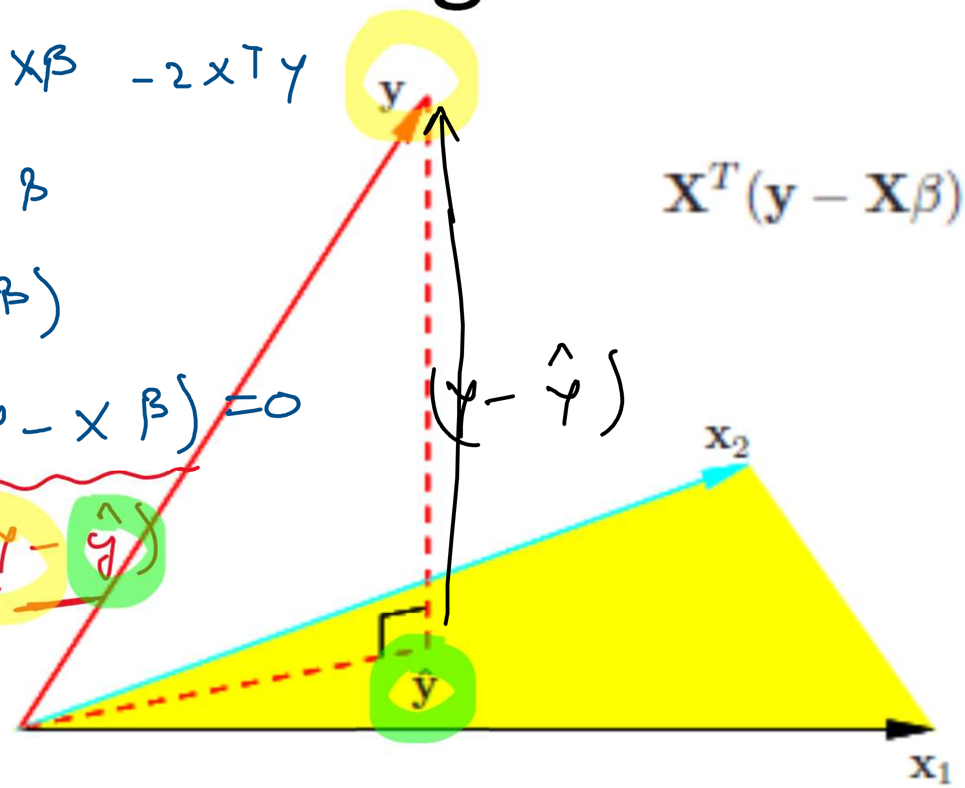
$= X^T y - X^T X \beta$

$= X^T (y - X\beta)$

for β^* $\quad X^T (y - X\beta) = 0$

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

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$



The N-dimensional geometry of least squares regression with two predictors. The outcome vector y is orthogonally projected onto the hyperplane spanned by the input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions

Gauss-Markov Theorem

- The Gauss–Markov theorem states that if we have any other linear estimator $\tilde{\theta} = \mathbf{c}^T \mathbf{y}$ that is unbiased for $\boldsymbol{\alpha}^T\beta$, that is, $E(c^T y) = \boldsymbol{\alpha}^T\beta$, then

$$\mathrm{Var}(a^T \hat{\beta}) \leq \mathrm{Var}(\mathbf{c}^T \mathbf{y}).$$

Gauss-Markov Theorem

- The least squares estimate of $\boldsymbol{\alpha}^T\beta$ is

$$\hat{\theta} = a^T\hat{\beta} = a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

- Considering X to be fixed, this is a linear function c_0^T y of the response vector y.
- If we assume that the linear model is correct, $\boldsymbol{\alpha}^T\beta$ is unbiased since

$$\begin{aligned}
\mathrm{E}(a^T\hat{\beta}) &= \mathrm{E}(a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\
&= a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta \\
&= a^T\beta.
\end{aligned}$$

Multiple Linear Regression

In multiple linear regression, we usually are interested in answering a few important questions.

1. *Is at least one of the predictors* $X1, X2, \ldots, Xp$ *useful in predicting the response?*

2. *Do all the predictors help to explain Y , or is only a subset of the predictors useful?*

3. *How well does the model fit the data?*

4. *Given a set of predictor values, what response value should we predict, and how accurate is our prediction?*

Is There a Relationship Between the Response and Predictors?

We test the null hypothesis,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

If this is true

$(TSS - RSS)/p \rightarrow \sigma^2$

versus the alternative

$$H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

This hypothesis test is performed by computing the F-statistic,

If H_0 is true

$F \approx 1$

If H_0 is not true

$F >> 1$

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n - p - 1)},$$

Linear model is correct

$\rightarrow \sigma^2$

Value of F-statistic close to 1 when null hypothesis true

Value of F-statistic greater than 1 when alternative hypothesis true

Hypothesis testing in multi linear regression

- F is very close to **one** we cannot reject the null hypothesis (thus, in a sense we accepted the **null hypothesis**)

- If F is **very large** we reject the null hypothesis (thus, in a sense we accepted the **alternate hypothesis**)

How large is large enough?

- This depends upon the values of n and p.
- If n is very large a small value above 1 is also a compelling evidence against the null hypothesis; however if n is a small then F has to be very large for us to reject the null hypothesis.

- When the null hypothesis is true and the error follows a Gaussian distribution, then it can be shown that F-statistic follows F-distribution

Hypothesis testing in multi linear regression

$$y = f(x) + \epsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots\cdots + \beta_p x_p + \epsilon$$

Simple linear regression

t-statistic $\Rightarrow$ p-value

$$t = \frac{\beta_1 - 0}{SE(\beta_1)}$$

For one parameter β_1

			t	p-value	H_0, H_a
x_1	β_1	$SE(\beta_1)$			
x_2	β_2	$SE(\beta_2)$			
$\vdots$	$\vdots$	$\vdots$			
x_p	β_p	$SE(\beta_p)$			

Hypothesis testing in multi linear regression

Why do we need F-statistic when t-statistic already exists?

(Given these individual p-values for each variable, why do we need to look at the overall F-statistic? After all, it seems likely that if any one of the p-values for the individual variables is very small, then at least one of the predictors is related to the response.)

- However, the above logic is flawed, especially when the number of predictors p is large.

- For instance, consider an example in which $p = 100$ and $H0 : \beta 1 = \beta 2 = . . . = \beta p = 0$ is true, so no variable is truly associated with the response. In this situation, about 5% of the p-values associated with each variable will be below 0.05 by chance. In other words, we expect to see approximately five small p-values even in the absence of any true association between the predictors and the response.

Extensions of Linear Models

- Removing the Additive Assumption

Introduce the interactive term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

- Where $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$

- Non-linear Relationships

$$\mathtt{mpg} = \beta_0 + \beta_1 \times \mathtt{horsepower} + \beta_2 \times \mathtt{horsepower}^2 + \epsilon$$

Nonlinear regression

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots$

$$\boxed{y = \beta_0 + \beta_1 x + \beta_2 x^2}$$

$\equiv y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$x \qquad x^2$

Find $\beta_0, \beta_1, \beta_2 \ldots$

Basis function regression

$$y = \beta_0 + \beta_1 f_1(\bar{X}) + \beta_2 f_2(\bar{x}) + \beta_3 f_3(\bar{X}) + \cdots + \epsilon$$

$$y = \underline{\beta_0} + \underline{\beta_1}\, f_1(x) + \underline{\beta_2}\, f_2(x) + \cdots + \epsilon$$

e.g. 1 $\qquad x \qquad x^2$

2. $\qquad f_1(x) = x^{1.5}, \quad f_2(x) = \left(1 + \frac{x}{3 + x^2}\right)$

Multiple Linear Regression

$y = \beta_0 + \beta_1 x + \epsilon$ $\rightarrow$ not continuous.

- Predictors with Only Two Levels
- Define new variables

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

The model then takes the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Multiple Linear Regression

- Predictors with more than Two Levels
- Define new variables

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

The model then takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American.} \end{cases}$$

Multiple Linear Regression

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

- Predictors with more than Two Levels
- Define new variables

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

y	x	x_1	x_2
y_1	AS	1	0
y_2	CA	0	1
y_3	AF	0	0
⋮	CA	⋮	⋮
⋮	AS	⋮	⋮
y_n	⋮	⋮	⋮

The model then takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American.} \end{cases}$$