STAT 420: Methods of Applied Statistics

Multiple Linear Regression

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

Course website: https://sites.google.com/site/teazrq/teaching/STAT420

University of Illinois at Urbana-Champaign
February 9, 2017

Multiple Linear Regression

- Usually a linear regression is perform a number of predictor:

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

- The techniques that we used earlier on simple linear regression can still be applied, but the calculation becomes very tedious.
- We have to setup $p + 1$ equations (taking derivatives of the SSE) and jointly solve for the optimizer.
- We are going to introduce a matrix representation of the solution that makes things easier.
- The distribution of the estimator will also be derived, which makes hypothesis testing possible.

Matrix representation

- The data that we have (from n such experiments) can be summarized into the following matrices:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}_{n\times(p+1)}$$

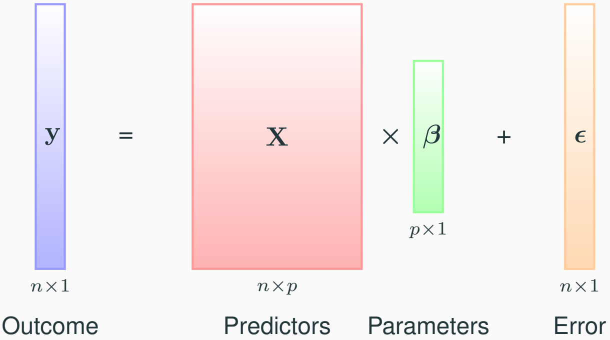
- The parameter vector $\boldsymbol{\beta}$ that we are interested has $p+1$ entries:

$$\boldsymbol{\beta}_{(p+1)\times 1} = (\beta_0, \beta_1, \ldots, \beta_p)^\mathsf{T}$$

- The linear regression can be represented as

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1}$$

Matrix representation



To clarify some notations

	Random Variable	Realization	Estimation
Outcome	Y	y	$\widehat{y}$, $\bar{y}$
Outcome of n samples	$\mathbf{Y}$	$\mathbf{y}$	$\widehat{\mathbf{y}}$
Predictor	$X, X_1, \ldots, X_p,$	x, x_i, x_{ij}	
Predictor of n samples		$\mathbf{X}$, $\mathbf{x}_j$	
Coefficients			$\widehat{\boldsymbol{\beta}}$
Error	ϵ		
Error of n samples	$\boldsymbol{\epsilon}$		$\mathbf{e}$

Matrix representation

- We can still calculate the sum of squared errors (SSE), based on any proposed β estimation

$$
\begin{aligned}
\text{SSE} &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}(y_i - x_i^\mathsf{T}\widehat{\boldsymbol{\beta}})^2 \\
&= \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2,
\end{aligned}
$$

  where x_i is the ith row of the design matrix $\mathbf{X}$, and $\|\cdot\|_2$ is called the ℓ_2-norm (Euclidean norm):

$$
\|\mathbf{a}\|_2 = \sqrt{\textstyle\sum_{i=1}^{n} a_i^2} = \sqrt{\mathbf{a}^\mathsf{T}\mathbf{a}}, \quad \text{and} \quad \|\mathbf{a}\|_2^2 = \textstyle\sum_{i=1}^{n} a_i^2 = \mathbf{a}^\mathsf{T}\mathbf{a}
$$

- We need to minimize the SSE

Matrix representation

- Again, we take derivative of the SSE and obtain a $p+1$ dimensional vector

$$\frac{\partial \text{SSE}}{\partial \boldsymbol{\beta}} = 2\sum_{i=1}^{n} x_i (y_i - x_i^\mathsf{T} \widehat{\boldsymbol{\beta}})$$
$$= 2\left(\mathbf{X}^\mathsf{T}\mathbf{y} - \mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol{\beta}}\right).$$

- Setting the above to be 0, we have $p+1$ equations represented in the matrix form:

$$\mathbf{X}^\mathsf{T}\mathbf{y} = \mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol{\beta}},$$

  which is called the normal equations.
- Validate that this is exactly the equations we had for the simple linear regression ($p = 1$ case). What is the design matrix $\mathbf{X}$?
- How to solve this?

Matrix representation

- In most of the cases $\mathbf{X}^\intercal\mathbf{X}$ is a positive definite matrix, this means we can multiple $(\mathbf{X}^\intercal\mathbf{X})^{-1}$ on both sides of the normal equations and obtain

$$\begin{aligned} & (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X}\widehat{\boldsymbol{\beta}} \\ \Longrightarrow \quad & (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} = \widehat{\boldsymbol{\beta}} \end{aligned}$$

  which gives us the solution.
- Why $\mathbf{X}^\intercal\mathbf{X}$ is usually positive definite? What if it is not? — The column vectors of $\mathbf{X}$ will be linearly dependent. This causes trouble...

Example

ID	Intercept	X_1	X_2	Y
1	1	0	1	11
2	1	11	5	15
3	1	11	4	13
4	1	7	3	14
5	1	4	1	0
6	1	10	4	19
7	1	5	4	16
8	1	8	2	8

- Setup the design matrix and response vector
- Perform MLR using solutions to the normal equation.

Example 1

```
> # set up the design matrix:
> X1 = c(0, 11, 11, 7, 4, 10, 5, 8)
> X2 = c(1, 5, 4, 3, 1, 4, 4, 2)
>
> X = cbind("Intercept" = 1, X1, X2)
>
> y = as.matrix(c(11, 15, 13, 14, 0, 19, 16, 8))

> # the final solution of beta
> solve(t(X) %*% X) %*% t(X) %*% y
           [,1]
Intercept   3.7
X1         -0.7
X2          4.4
```

Example 1

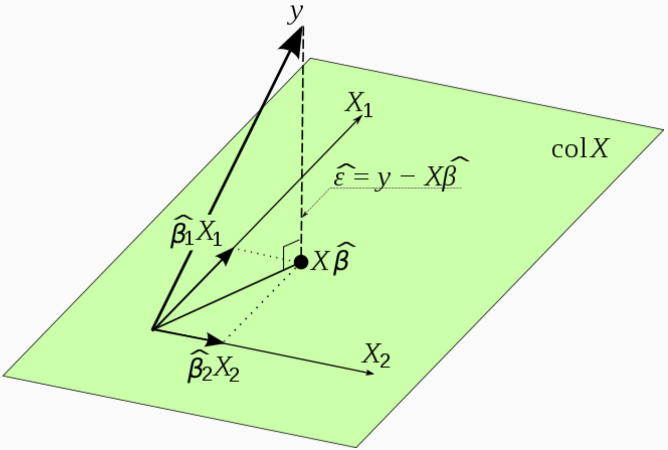
```
> # check it with lm(), by default, the intercept term will be
    included
>
> lm(y ~ X1 + X2)

Call:
lm(formula = y ~ X1 + X2)

Coefficients:
(Intercept)           X1           X2
        3.7         -0.7          4.4
```

Geometric interpretation

- Linear regression can be viewed as projecting the vector $\mathbf{y}$ onto a hyperplane defined by the column space of $\mathbf{X}$



Geometric interpretation

- The column vectors of $\mathbf{X}$ are

$$\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix}, \quad \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix}, \quad \cdots \quad \begin{pmatrix} x_{np} \\ x_{np} \\ \vdots \\ x_{np} \end{pmatrix}$$

- Any element in the column space col$(\mathbf{X})$ of $\mathbf{X}$ can be expressed as their linear combinations:

$$\beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} + \beta_2 \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix} + \cdots + \beta_p \begin{pmatrix} x_{np} \\ x_{np} \\ \vdots \\ x_{np} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta}$$

Geometric interpretation

- Among all these kind of linear combinations (search through the entire column space of $\mathbf{X}$, namely col$(\mathbf{X})$), find the one closest to $\mathbf{y}$.
- How to define "closest"? — Euclidean distance, the ℓ_2 norm.
- This is the same as projecting the vector $\mathbf{y}$ onto the space col$(\mathbf{X})$ (shown in the previous plot).
- The projection is $\widehat{\mathbf{y}}$, and the remaining part $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}$ will be orthogonal to the space col$(\mathbf{X})$.
- There are some easy ways to calculate this project.

Special case: orthogonal design matrix

- Usually it is difficult to calculate the inverse matrix $(\mathbf{X}^\mathsf{T}\mathbf{X})$, however, there is a special case when $\mathbf{X}^\mathsf{T}\mathbf{X}$ is an diagonal matrix, i.e., only the diagonal elements are non-zero.
- This happens when the columns of $\mathbf{X}$ are orthogonal to each other.
- An example:

Intercept	X_1	X_2	Y
1	1	1	1
1	1	-1	2
1	-1	1	3
1	-1	-1	4

- Calculate the regression coefficients by hand.

Hand calculation of the $\widehat{\beta}$

- We first get $\mathbf{X}^{\mathsf{T}}\mathbf{X}$, which is a diagonal matrix

$$\mathbf{X}^{\mathsf{T}}\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{pmatrix} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$$

- The inverse of that is just taking the inverse of each element:

$$(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} = \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{pmatrix}$$

- Multiple that to the $\mathbf{X}^{\mathsf{T}}\mathbf{y}$, we have

$$(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} = \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{pmatrix} \begin{pmatrix} 10 \\ -4 \\ -2 \end{pmatrix} = \begin{pmatrix} 2.5 \\ -1 \\ -0.5 \end{pmatrix}$$

- However, you can check that this is a perfect fit, meaning that $\widehat{\mathbf{y}} = \mathbf{y}$ exactly, which not good...

Geometric interpretation

- Let $\mathbf{H} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$ be a projection matrix referred to as the "hat" matrix

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$$
$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

- $\mathbf{H}$ is idempotent: $\mathbf{H}$ is symmetric and $\mathbf{H}\mathbf{H} = \mathbf{H}$

Proof.

$$\begin{aligned}
\mathbf{H}^\mathsf{T} &= \left(\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\right)^\mathsf{T} \\
&= \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} = \mathbf{H} \\
\mathbf{H}\mathbf{H} &= \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} \\
&= \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} = \mathbf{H}
\end{aligned}$$

□

Example

- Load the cheddar data, model taste using all three covaraites (with intercept): Acetic, H2S and Lactic.
- Calculate the following quantities to perform MLR:
  - $\mathbf{X}^\top\mathbf{X}$, and check if it is positive definite
  - The parameter estimates $\widehat{\boldsymbol{\beta}}$
  - $\mathbf{H}$, calculate SSE and $\widehat{\sigma}^2$, what is the degrees of freedom?
  - The coefficient of determination R^2

Example

- Load the `cheddar` data, model `taste` using all three covaraites (with intercept): Acetic, H2S and Lactic.
- Calculate the following quantities to perform MLR:
  - $\mathbf{X}^\top \mathbf{X}$, and check if it is positive definite
  - The parameter estimates $\widehat{\boldsymbol{\beta}}$
  - $\mathbf{H}$, calculate SSE and $\widehat{\sigma}^2$, what is the degrees of freedom?
  - The coefficient of determination R^2
- If a researcher wants to use at most 2 covariates, which is the best model?

A: Acetic + H2S; B: H2S + Lactic; C: Acetic + Lactic

Example

- Load the cheddar data, model taste using all three covaraites (with intercept): Acetic, H2S and Lactic.
- Calculate the following quantities to perform MLR:
  - $\mathbf{X}^\top\mathbf{X}$, and check if it is positive definite
  - The parameter estimates $\widehat{\boldsymbol{\beta}}$
  - $\mathbf{H}$, calculate SSE and $\widehat{\sigma}^2$, what is the degrees of freedom?
  - The coefficient of determination R^2
- If a researcher wants to use at most 2 covariates, which is the best model?

  A: Acetic + H2S; B: H2S + Lactic; C: Acetic + Lactic

- Question: Will MLR with X_1 and X_2 always outperforms the model using X_1 only?

The sum of squares

- Recall that SST $=$ SSR $+$ SSE

$$
\begin{aligned}
\text{SST} &= \|\mathbf{y} - \bar{y}\mathbf{1}\|_2^2 \\
\text{SSE} &= \|(\mathbf{I} - \mathbf{H})\mathbf{y}\|_2^2 = \mathbf{y}^\mathsf{T}(\mathbf{I} - \mathbf{H})^\mathsf{T}(\mathbf{I} - \mathbf{H})\mathbf{y} \\
&= \mathbf{y}^\mathsf{T}(\mathbf{I} - \mathbf{H})\mathbf{y} \\
\text{SSR} &= \|\mathbf{X}\widehat{\boldsymbol{\beta}} - \bar{y}\mathbf{1}\|_2^2
\end{aligned}
$$

- $\mathbf{1}$ is a vector of length n, with each element being 1.
- SST resides in $n-1$ dimensions; SSE in $n-p-1$ dimensions; SSR in p dimensions.
- Careful: Sometimes people count the intercept as one of the p dimensions, we didn't.