Linear Regression II

Nipun Batra and the teaching staff January 20, 2020

IIT Gandhinagar

If N< M, then it is an under-determined system

If N< M, then it is an under-determined system Example: N=2; M=3

If N< M, then it is an under-determined system Example: N=2; M=3

$$\begin{bmatrix} 30\\ 40 \end{bmatrix} = \begin{bmatrix} 1 & 6 & 30\\ 1 & 5 & 20 \end{bmatrix} \begin{bmatrix} \theta_0\\ \theta_1\\ \theta_2 \end{bmatrix}$$

If N< M, then it is an under-determined system Example: N=2; M=3

$$\begin{bmatrix} 30\\40 \end{bmatrix} = \begin{bmatrix} 1 & 6 & 30\\1 & 5 & 20 \end{bmatrix} \begin{bmatrix} \theta_0\\\theta_1\\\theta_2 \end{bmatrix}$$

$$30 = \theta_0 + 6\theta_1 + 30\theta_2$$

$$40 = \theta_0 + 5\theta_1 + 20\theta_2$$

$$-10 = -1\theta_1 - 10\theta_2$$
(1)

The above equation can have infinitely many solutions. Under-determined system: $\epsilon_i = 0$ for all *i*

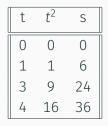
What if N > M

What if N > MThen it is an over determined system. So, the sum of squared residuals > 0. Transform the data, by including the higher power terms in the feature space.

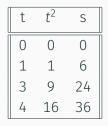
t	S
0	0
1	6
3	24
4	36

The above table represents the data before transformation

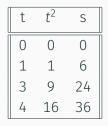
t	t ²	S
0	0	0
1	1	6
3	9	24
4	16	36



The above table represents the data after transformation



The above table represents the data after transformation Now, we can write $\hat{s} = f(t, t^2)$



The above table represents the data after transformation Now, we can write $\hat{s} = f(t, t^2)$ Other transformations: $\log(x), x_1 \times x_2$ 1. $\hat{s} = \theta_0 + \theta_1 * t$ is linear

¹https://stats.stackexchange.com/questions/8689/ what-does-linear-stand-for-in-linear-regression

¹https://stats.stackexchange.com/questions/8689/ what-does-linear-stand-for-in-linear-regression

1.
$$\hat{s} = \theta_0 + \theta_1 * t$$
 is linear
2. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$ linear?
3. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2 + \theta_3 * \cos(t^3)$ linear?

¹https://stats.stackexchange.com/questions/8689/ what-does-linear-stand-for-in-linear-regression

1.
$$\hat{s} = \theta_0 + \theta_1 * t$$
 is linear
2. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$ linear?
3. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2 + \theta_3 * \cos(t^3)$ linear?
4. Is $\hat{s} = \theta_0 + \theta_1 * t + e^{\theta_2} * t$ linear?

¹https://stats.stackexchange.com/questions/8689/ what-does-linear-stand-for-in-linear-regression

1.
$$\hat{s} = \theta_0 + \theta_1 * t$$
 is linear
2. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$ linear?
3. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2 + \theta_3 * \cos(t^3)$ linear?
4. Is $\hat{s} = \theta_0 + \theta_1 * t + e^{\theta_2} * t$ linear?

5. All except #4 are linear models!

¹https://stats.stackexchange.com/questions/8689/ what-does-linear-stand-for-in-linear-regression

- 1. $\hat{s} = \theta_0 + \theta_1 * t$ is linear
- 2. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$ linear?
- 3. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2 + \theta_3 * \cos(t^3)$ linear?
- 4. Is $\hat{s} = \theta_0 + \theta_1 * t + e^{\theta_2} * t$ linear?
- 5. All except #4 are linear models!
- 6. Linear refers to the relationship between the parameters that you are estimating (θ) and the outcome

¹https://stats.stackexchange.com/questions/8689/ what-does-linear-stand-for-in-linear-regression

Solve the linear system below using normal equation method

<i>X</i> ₁	<i>X</i> ₂	У
1	2	4
2	4	6
3	6	8

There can be situations where $X^T X$ is not computable.

There can be situations where $X^T X$ is not computable. This condition arises when the $|X^T X| = 0$.

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{bmatrix}$$

(2)

There can be situations where $X^T X$ is not computable. This condition arises when the $|X^T X| = 0$.

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{bmatrix}$$

The matrix X is not full rank.

(2)

It arises when one or more predictor varibale/feature in X can be expressed as a linear combinations of others

How to tackle it?

• Regularize

It arises when one or more predictor varibale/feature in X can be expressed as a linear combinations of others

How to tackle it?

- Regularize
- Drop variables

It arises when one or more predictor varibale/feature in X can be expressed as a linear combinations of others

How to tackle it?

- Regularize
- Drop variables
- Use different subsets of data

It arises when one or more predictor varibale/feature in X can be expressed as a linear combinations of others

How to tackle it?

- Regularize
- Drop variables
- Use different subsets of data
- Avoid dummy variable trap

 $P = \theta_0 + \theta_1 # Wehicles + \theta_1 Wind speed + \theta_3 Wind Direction$

 $P = \theta_0 + \theta_1^*$ #Vehicles + θ_1^* Wind speed + θ_3^* Wind Direction

But, wind direction is a categorical variable.

 $P = \theta_0 + \theta_1^*$ #Vehicles + θ_1^* Wind speed + θ_3^* Wind Direction

But, wind direction is a categorical variable. It is denoted as follows {N:0, E:1, W:2, S:3 }

 $P = \theta_0 + \theta_1^*$ #Vehicles + θ_1^* Wind speed + θ_3^* Wind Direction

But, wind direction is a categorical variable. It is denoted as follows {N:0, E:1, W:2, S:3 }

Can we use the direct encoding?

 $P = \theta_0 + \theta_1^*$ #Vehicles + θ_1^* Wind speed + θ_3^* Wind Direction

But, wind direction is a categorical variable. It is denoted as follows {N:0, E:1, W:2, S:3 }

Can we use the direct encoding? Then this implies that S>W>E>N

N-1 Variable encoding

	Is it N?	Is it E?	Is it W?
Ν	1	0	0
E	0	1	0
W	0	0	1
S	0	0	0

N Variable encoding

	Is it N?	Is it E?	Is it W?	Is it S?
Ν	1	0	0	0
E	0	1	0	0
W	0	0	1	0
S	0	0	0	1

Which is better N variable encoding or N-1 variable encoding?

Which is better N variable encoding or N-1 variable encoding? The N-1 variable encoding is better because the N variable encoding can cause multi-collinearity. Which is better N variable encoding or N-1 variable encoding? The N-1 variable encoding is better because the N variable encoding can cause multi-collinearity. Is it S = 1 - (Is it N + Is it W + Is it E)

Ν	00
E	01
W	10
S	11

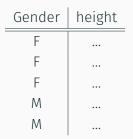
Ν	00
E	01
W	10
S	11

W and S are related by one bit.

Ν	00
E	01
W	10
S	11

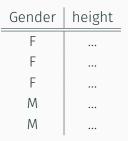
W and S are related by one bit.

This introduces dependencies between them, and this can confusion in classifiers.

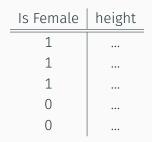




Encoding



Encoding



14

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

 $height_i = \theta_0 + \theta_1 * (Is Female) + \epsilon_i$

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

 $height_i = \theta_0 + \theta_1 * (Is Female) + \epsilon_i$

We get θ_0 = 5.8 and θ_0 = 6

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

 $height_i = \theta_0 + \theta_1 * (Is Female) + \epsilon_i$

We get θ_0 = 5.8 and θ_0 = 6 θ_0 = Avg height of Male = 5.9

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

 $height_i = \theta_0 + \theta_1 * (Is Female) + \epsilon_i$

We get $\theta_0 = 5.8$ and $\theta_0 = 6$ $\theta_0 = Avg$ height of Male = 5.9 $\theta_0 + \theta_1$ is chosen based (equal to) on 5, 5.2, 5.4 (for three records).

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

 $height_i = \theta_0 + \theta_1 * (Is Female) + \epsilon_i$

We get θ_0 = 5.8 and θ_0 = 6 θ_0 = Avg height of Male = 5.9 $\theta_0 + \theta_1$ is chosen based (equal to) on 5, 5.2, 5.4 (for three records).

 θ_1 is chosen based on 5-5.9, 5.2-5.9, 5.4-5.9

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

 $height_i = \theta_0 + \theta_1 * (Is Female) + \epsilon_i$

We get θ_0 = 5.8 and θ_0 = 6 θ_0 = Avg height of Male = 5.9 $\theta_0 + \theta_1$ is chosen based (equal to) on 5, 5.2, 5.4 (for three records).

*θ*₁ is chosen based on 5-5.9, 5.2-5.9, 5.4-5.9 *θ*₁ = Avg. female height (5+5.2+5.4)/3 - Avg. male height(5.9)

$$\hat{y}_i = \theta_0 + \theta_1 x_i$$

$$\hat{y}_i = \theta_0 + \theta_1 x_i$$

$$\epsilon_i = y_i - \hat{y}_i$$

$$\hat{y}_i = \theta_0 + \theta_1 x_i$$

$$\epsilon_i = y_i - \hat{y}_i$$

$$\sum \epsilon_i^2 = \sum (y_i - \theta_0 - \theta_1 x_i)^2$$

$$\hat{y}_i = \theta_0 + \theta_1 x_i$$

$$\epsilon_i = y_i - \hat{y}_i$$

$$\sum \epsilon_i^2 = \sum (y_i - \theta_0 - \theta_1 x_i)^2$$

Now, we compute the derivative of it with all the θ_j . Let us solve for x being a scalar.

$$\frac{\partial}{\partial \theta_0} \sum \epsilon_i^2 = 2 \sum (y_i - \theta_0 - \theta_1 x_i)(-1) = 0$$

$$0 = \sum y_i - N\theta_0 - \sum \theta_1 x_i$$

$$\theta_0 = \frac{\sum y_i - \theta_1 \sum x_i}{N}$$
(3)

$$heta_0=ar{y}- heta_1ar{x}$$

$$\frac{\partial}{\partial \theta_1} \sum \epsilon_i^2 = 0$$

$$\implies 2\sum_{i=1}^{N}(y_i-\theta_0-\theta_1x_i)(-x_i)=0$$

$$\implies \sum_{i=1}^{N} (x_i y_i - \theta_0 x_i - \theta_1 x_i^2) = 0$$

$$\implies \sum \theta_1 x_i^2 = \sum x_i y_i - \sum \theta_0 x_i$$

$$\implies \sum \theta_1 x_i^2 = \sum x_i y_i - \sum (\bar{y} - \theta_1 \bar{x}) x_i$$

$$\implies \sum \theta_1 x_i^2 = \sum x_i y_i - \bar{y} \sum x_i + \theta_1 \bar{x} \sum x_i$$
$$\implies \sum x_i y_i - \sum x_i y = \theta_1 (-\bar{x} \sum x_i + \sum x_i^2)$$
$$a = x_i y_i - \sum x_i y$$

$$\theta_1 = \frac{x_i y_i - \sum x_i y}{\sum x_i^2 - \bar{x} \sum x_i}$$

$$\theta_{1} = \frac{\frac{1}{N} \sum_{i=1}^{N} (x_{i} - \bar{x})(y_{i} - \bar{y})}{\frac{1}{N} (x_{i} - \bar{x})^{2}}$$

$$\theta_1 = \frac{Cov(x, y)}{variance(x)}$$