Linear Regression II

Nipun Batra and the teaching staff

January 20, 2020

IIT Gandhinagar

Relation between #instances and # Variables

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

Relation between #instances and # Variables

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

Example: N=2; M=3

Relation between #instances and # Variables

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

Example: N=2; M=3

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

Relation between #instances and # Variables

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

Example: N=2; M=3

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

$$\begin{array}{r} 30 = \theta_0 + 6\theta_1 + 30\theta_2 \\ 40 = \theta_0 + 5\theta_1 + 20\theta_2 \\ \hline -10 = -1\theta_1 - 10\theta_2 \end{array} \qquad (1)$$

The above equation can have infinitely many solutions.

Under-determined system: $\epsilon_i = 0$ for all i

Relation between #instances and # Variables

What if $N > M$

Relation between #instances and # Variables

What if $N > M$

Then it is an over determined system. So, the sum of squared residuals > 0.

Variable Transformation

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

Variable Transformation

Add the higher degree features to the previous table

t	t^2	s
0	0	0
1	1	6
3	9	24
4	16	36

Variable Transformation

Add the higher degree features to the previous table

t	t^2	s
0	0	0
1	1	6
3	9	24
4	16	36

The above table represents the data after transformation

Variable Transformation

Add the higher degree features to the previous table

t	t^2	s
0	0	0
1	1	6
3	9	24
4	16	36

The above table represents the data after transformation

Now, we can write $\hat{s} = f(t, t^2)$

Variable Transformation

Add the higher degree features to the previous table

t	t^2	s
0	0	0
1	1	6
3	9	24
4	16	36

The above table represents the data after transformation

Now, we can write $\hat{s} = f(t, t^2)$

Other transformations: $\log(x), x_1 \times x_2$

A big caveat: Linear in what?![1]

1. $\hat{s} = \theta_0 + \theta_1 * t$ is linear

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

A big caveat: Linear in what?![1]

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

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

A big caveat: Linear in what?![1]

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?

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

A big caveat: Linear in what?![1]

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?

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

A big caveat: Linear in what?![1]

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 $\hat{s} = \theta_0 + \theta_1 * t + \mathrm{e}^{\theta_2} * t$ linear?
5. All except #4 are linear models!

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

A big caveat: Linear in what?![1]

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 $\hat{s} = \theta_0 + \theta_1 * t + \mathrm{e}^{\theta_2} * t$ 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

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

Class Exercise

Solve the linear system below using normal equation method

x_1	x_2	y
1	2	4
2	4	6
3	6	8

Multi-collinearity

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

Multi-collinearity

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

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

Multi-collinearity

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

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

The matrix X is not full rank.

Multi-collinearity

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

Multi-collinearity

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

Multi-collinearity

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

Multi-collinearity

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

Dummy variables

Say Pollution in Delhi = P

Dummy variables

Say Pollution in Delhi = P

P = θ_0 + θ_1*#Vehicles + θ_1* *Wind speed* + θ_3 * *Wind Direction*

Dummy variables

Say Pollution in Delhi = P

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

But, wind direction is a categorical variable.

Dummy variables

Say Pollution in Delhi = P

$P = \theta_0 + \theta_1$*#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 }

Dummy variables

Say Pollution in Delhi = P

$P = \theta_0 + \theta_1$*#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?

Dummy variables

Say Pollution in Delhi = P

P = θ_0 + θ_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?
N	1	0	0
E	0	1	0
W	0	0	1
S	0	0	0

Dummy Variables

N Variable encoding

	Is it N?	Is it E?	Is it W?	Is it S?
N	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?

Dummy Variables

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.

Dummy Variables

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)

Binary Encoding

N	00
E	01
W	10
S	11

Binary Encoding

N	00
E	01
W	10
S	11

W and S are related by one bit.

Binary Encoding

N	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.

Interpreting Dummy variables

Gender	height
F	...
F	...
F	...
M	...
M	...

Interpreting Dummy variables

Gender	height
F	...
F	...
F	...
M	...
M	...

Encoding

Interpreting Dummy variables

Gender	height
F	...
F	...
F	...
M	...
M	...

Encoding

Is Female	height
1	...
1	...
1	...
0	...
0	...

Interpreting Dummy Variables

Interpreting Dummy Variables

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

Interpreting Dummy Variables

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

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

Interpreting Dummy Variables

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

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

We get $\theta_0 = 5.8$ and $\theta_0 = 6$

Interpreting Dummy Variables

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

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

We get $\theta_0 = 5.8$ and $\theta_0 = 6$

θ_0 = Avg height of Male = 5.9

Interpreting Dummy Variables

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

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

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).

Interpreting Dummy Variables

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

$height_i$ = θ_0 + θ_1 * (Is Female) + ϵ_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

Interpreting Dummy Variables

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

$height_i$ = θ_0 + θ_1 * (Is Female) + ϵ_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)

Alternative parameter estimation

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

Alternative parameter estimation

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

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

Alternative parameter estimation

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

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

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

Alternative parameter estimation

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

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

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

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

Alternative parameter estimation

$$\frac{\partial}{\partial \theta_0} \sum \epsilon_i^2 = 2 \sum (y_i - \theta_0 - \theta_1 x_i)(-1) = 0$$
$$0 = \sum y_i - N\theta_0 - \sum \theta_1 x_i \tag{3}$$
$$\theta_0 = \frac{\sum y_i - \theta_1 \sum x_i}{N}$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

Alternative parameter estimation

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

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

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

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

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

Alternative parameter estimation

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

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

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

Alternative parameter estimation

$$\theta_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N}(x_i - \bar{x})^2}$$

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