Math 555E

Atabey Kaygun

Thursday, April 1, 2021

Regression Models

The Basic Setup

We have **structured** data (column data)

```
     CIC0   SM1Dz  GATS1i MLOGP  LC50
0    3.260  0.829  1.676  1.453  3.770
1    2.189  0.580  0.863  1.348  3.115
2    2.125  0.638  0.831  1.348  3.531
...  ...  ...  ...  ...  ...  ...
905 3.763  0.916  0.878  2.918  4.818
906 2.831  1.393  1.077  0.906  5.317
907 4.057  1.032  1.183  4.754  8.201
```

- Predict the last column

The Basic Setup

We have **structured** data (column data)

```
     CIC0   SM1Dz  GATS1i  MLOGP  LC50
0    3.260  0.829  1.676   1.453  3.770
1    2.189  0.580  0.863   1.348  3.115
2    2.125  0.638  0.831   1.348  3.531
...  ...  ...  ...  ...  ...  ...
905  3.763  0.916  0.878   2.918  4.818
906  2.831  1.393  1.077   0.906  5.317
907  4.057  1.032  1.183   4.754  8.201
```

- Predict the last column
- We form a linear model

$$\text{LC50} \approx \alpha + \beta_1\ \text{CIC0} + \beta_2\text{SM1Dz} + \beta_3\text{GATS1i} + \beta_4\text{MLOGP}$$

The Optimization Model

The **best-fitting** linear model

$$\mathrm{argmin}_{\alpha,\beta}\|LC50 - \alpha - \beta \cdot X\|$$

The Questions

- Did the model fit?

The Questions

- Did the model fit?
- How well?

The Questions

- Did the model fit?
- How well?
- Which variables are (more) important?

A Common Misconception

The size of the coefficient

- doesn't mean importance

A Common Misconception

The size of the coefficient

- doesn't mean importance
- the coefficient relates to relative size

Coefficient of Determinatation (R^2)

- Assume we have a linear model

$$Y \approx \beta \cdot X$$

Coefficient of Determinatation (R^2)

- Assume we have a linear model

$$Y \approx \beta \cdot X$$

- Total variance in the dependent variable

$$SS_{tot} = \frac{1}{N} \sum_i (Y_i - \hat{Y})^2$$

Coefficient of Determinatation (R^2)

- Assume we have a linear model

$$Y \approx \beta \cdot X$$

- Total variance in the dependent variable

$$SS_{tot} = \frac{1}{N} \sum_i (Y_i - \hat{Y})^2$$

- Total residual variance

$$SS_{res} = \frac{1}{N} \sum_i (Y_i - \beta \cdot X_i)^2$$

Coefficient of Determinatation (R^2)

- Total **explained** variance

$$\frac{SS_{res}}{SS_{tot}}$$

Coefficient of Determinatation (R^2)

- Total **explained** variance

$$\frac{SS_{res}}{SS_{tot}}$$

- Total **unexplained** variance

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

ANOVA (ANalysis Of VAriance)

- Instead of **total** explained variance do one variable

$$1 - \frac{SS_{res,i}}{SS_{tot}}$$

where

$$SS_{res,i} = \frac{1}{N}\sum_i (Y_i - \beta_j X_{ij})^2$$

DEMO

Variable Types

Numerical vs Categorical

- A numerical variable takes values in $\mathbb{R}$

Numerical vs Categorical

- A numerical variable takes values in $\mathbb{R}$
- A categorical variable is discrete.

Numerical vs Categorical

- A numerical variable takes values in $\mathbb{R}$
- A categorical variable is discrete.
  - Ordered

Numerical vs Categorical

- A numerical variable takes values in $\mathbb{R}$
- A categorical variable is discrete.
  - Ordered
  - Unordered

Numerical vs Categorical

- A numerical variable takes values in $\mathbb{R}$
- A categorical variable is discrete.
  - Ordered
  - Unordered
- Examples

Numerical vs Categorical

- A numerical variable takes values in $\mathbb{R}$
- A categorical variable is discrete.
  - Ordered
  - Unordered
- Examples
  - Education levels: primary $<$ secondary $<$ college $<$ MSc $<$ PhD

Numerical vs Categorical

- A numerical variable takes values in $\mathbb{R}$
- A categorical variable is discrete.
  - Ordered
  - Unordered
- Examples
  - Education levels: primary $<$ secondary $<$ college $<$ MSc $<$ PhD
  - Socio-Economic level: low $<$ middle $<$ high

Numerical vs Categorical

- A numerical variable takes values in $\mathbb{R}$
- A categorical variable is discrete.
  - Ordered
  - Unordered
- Examples
  - Education levels: primary $<$ secondary $<$ college $<$ MSc $<$ PhD
  - Socio-Economic level: low $<$ middle $<$ high
  - Car brands: Toyota, Mercedes, BMW, Fiat, ...

Problems with Categorical Variables

- Do not automatically encode them with numbers

Problems with Categorical Variables

- Do not automatically encode them with numbers
  - Class: low $=0$, middle $=1$, high $=2$

Problems with Categorical Variables

- Do not automatically encode them with numbers
  - Class: low $= 0$, middle $= 1$, high $= 2$
  - Would it mean

Problems with Categorical Variables

- Do not automatically encode them with numbers
  - Class: low $= 0$, middle $= 1$, high $= 2$
  - Would it mean
    - the difference btw low and middle is the same as middle and high?

Problems with Categorical Variables

- Do not automatically encode them with numbers
  - Class: low $= 0$, middle $= 1$, high $= 2$
  - Would it mean
    - the difference btw low and middle is the same as middle and high?
    - the difference btw low and high is twice as much btw low and middle?

Problems with Categorical Variables

- Do not automatically encode them with numbers
  - Class: low $=0$, middle $=1$, high $=2$
  - Would it mean
    - the difference btw low and middle is the same as middle and high?
    - the difference btw low and high is twice as much btw low and middle?
  - In the case variable instances are not ordered

Problems with Categorical Variables

- Do not automatically encode them with numbers
  - Class: low $= 0$, middle $= 1$, high $= 2$
  - Would it mean
    - the difference btw low and middle is the same as middle and high?
    - the difference btw low and high is twice as much btw low and middle?
  - In the case variable instances are not ordered
    - Toyota $= 0$, Mercedes $= 1$, BMW $= 2$

Problems with Categorical Variables

- Do not automatically encode them with numbers
    - Class: low $= 0$, middle $= 1$, high $= 2$
    - Would it mean
        - the difference btw low and middle is the same as middle and high?
        - the difference btw low and high is twice as much btw low and middle?
    - In the case variable instances are not ordered
        - Toyota $= 0$, Mercedes $= 1$, BMW $= 2$
        - Toyota $<$ Mercedes $<$ BMW?

Encodings of Categorical Variables

Examples

- One Hot Encoding

Encodings of Categorical Variables

Examples

- One Hot Encoding
- Label/Ordinal Encoding

Encodings of Categorical Variables

Examples

- One Hot Encoding
- Label/Ordinal Encoding
- Frequency Encoding

Encodings of Categorical Variables

Examples

- One Hot Encoding
- Label/Ordinal Encoding
- Frequency Encoding
- Mean Encoding

Encodings of Categorical Variables

Examples

- One Hot Encoding
- Label/Ordinal Encoding
- Frequency Encoding
- Mean Encoding
- Hash-Encoding

Regression with Categorical Variables

Demo