ECS 256 – Project Bias, Variance and Parsimony in Regression Analysis

Xinbo Wang, Divya Chitimalla, Abhishek Roy, Aveek Das

Bias Calculation for Linear Model

- Objective of Linear Regression is to minimize the mean square error

$$E = \text{Mean Squared Error (MSE)} = \frac{1}{N}\sum_{i=1}^{N}(m_{Y;X}(t_i) - \hat{m}_{Y;X}(t_i))^2$$
$$= \frac{1}{N}\sum_{i=1}^{N}(m_{Y;X}(t_i) - \beta t_i)^2$$

- For the optimal estimate of slope we take the derivate of the error with respect to slope and equate to zero

$$\frac{\partial E}{\partial \beta} = 0$$

Bias Calculation for Linear Model

Doing the calculus we obtain the slope as

$$\Rightarrow \beta = \frac{\frac{1}{N}\sum_{i=1}^{N} t_i \cdot m_{Y;X}(t_i)}{\frac{1}{N}\sum_{i=1}^{N} t_i^2}$$

For a very large value of N we have

$$N \to \infty$$

$$\beta \to \frac{E(t m_{Y;X}(t))}{E(t^2)} = \frac{E(t^{1.75})}{E(t^2)}$$

2a – *dimension reduction* Parsimony

- Problem with Significance Testing – Everything is significant in Big Data Sets
- Prediction accuracy criterion (PAC) 1–k fits your definition of "almost."
- Adjusted R^2 – Another metric to decide the accuracy of the reduced model

Functions and Structure

Main function that takes in the full model, PAC value, Model type to output the parsimonious model

- prsm(y,x,k=0.01,predacc=ar2,crit=NULL,printdel=F)

Function to return summary of generalized linear model

- aiclogit (y,x)

Function to return summary of linear model

- ar2 (y,x)

Function to return the reduced data set

- findRes (index, nmax)

Results for 2a using Diabetics Data

When using linear model
full outcome = 0.2959093
deleted Thick
new outcome = 0.2968178
deleted Insul
new outcome = 0.2962828
The variables used in this model are: NPreg Gluc BP BMI Genet Age

When using generalized linear model
full outcome = 741.4454
deleted Thick
new outcome = 739.4534
deleted Insul
new outcome = 739.4617
deleted BP
new outcome = 744.5088
deleted Age
new outcome = 744.3059
The variables used in this model are: NPreg Gluc BMI Genet

2b – Simulation using known distribution

Let $X_1,\ldots,X_{10}$ be i.i.d. $U(0,1)$, with

$m_X(t) = t_1 + t_2 + t_3 + 0.1\ t_4 + 0.01\ t_5$
and with the distribution of Y given X being

$U(m-1,m+1)$, where m means m_X

2b – Simulation Results

When n = 100, k = 0.01
First run : The variables used in this model are: x1 x2 x3 x4 x10
Second run:The variables used in this model are: x1 x2 x3 x5 x6 x8
Third run: The variables used in this model are: x1 x2 x3 x5 x6

When n = 100, k = 0.05
First run : The variables used in this model are: x1 x2 x3
Second run:The variables used in this model are: x1 x2 x3
Third run : The variables used in this model are: x1 x2 x3

when n = 1000, k = 0.01
First run : The variables used in this model are: x1 x2 x3 x6 x8 x10
Second run: The variables used in this model are: x1 x2 x3 x5
Third run : The variables used in this model are: x1 x2 x3 x4

2b – Functions and Structure

- Function to test the model using simulation test(n,k)

- Function to calculate the known distribution

calY(x)

2b – Simulation Results

when n = 1000, k = 0.05
first run : The variables used in this model are: x1 x2 x3
Second run: The variables used in this model are: x1 x2 x3
Third run :The variables used in this model are: x1 x2 x3

when n = 10000, k = 0.01
first run : The variables used in this model are: x1 x2 x3 x10
Second run: The variables used in this model are: x1 x2 x3 x8 x9
Third run : The variables used in this model are: x1 x2 x3 x4 x6

when n = 10000, k = 0.05
first run : The variables used in this model are: x1 x2 x3
Second run The variables used in this model are: x1 x2 x3
Third run The variables used in this model are: x1 x2 x3

2b – Simulation Results

when n = 100000, k = 0.01
first run : The variables used in this model are: x1 x2 x3 x10
Second run: The variables used in this model are: x1 x2 x3 x5 x9
Third run: The variables used in this model are: x1 x2 x3 x5

when n = 100000, k = 0.05
first run : The variables used in this model are: x1 x2 x3
Second run: The variables used in this model are: x1 x2 x3
Third run: The variables used in this model are: x1 x2 x3

Results Using Significance Testing

Select predictors that is "significant" at the 5% level of less by running full model. (**bolded**) : x1, x2,x3,x9

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.46262	0.33979	1.362	0.176789	
x1	0.92421	0.22679	4.075	9.97e-05	***
x2	0.87121	0.21182	4.113	8.69e-05	***
x3	0.90259	0.22743	3.969	0.000146	***
x4	0.04334	0.21403	0.202	0.839992	
x5	0.03630	0.22842	0.159	0.874078	
x6	-0.09983	0.21858	-0.457	0.649004	
x7	-0.27588	0.22308	-1.237	0.219456	
x8	0.18937	0.22830	0.829	0.409062	
x9	-0.45749	0.21950	-2.084	0.040007	*
x10	0.11414	0.22266	0.513	0.609478	

2c – Discrete Case n<1000 p<10 0-1Y breast cancer Wisconsin

use 2~10 attributes to predict the 11th attribute: class

-https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/

Attribute Information: (class attribute has been moved to last column)   # Attribute   Name in dataset
Domain   1. Sample code number   Id   id number   2. Clump Thickness   Thick
1 - 10   3. Uniformity of Cell Size   Size   1 - 10   4. Uniformity of Cell Shape   Shape
1 - 10   5. Marginal Adhesion   Adh   1 - 10   6. Single Epithelial Cell Size   SECS
1 - 10   7. Bare Nuclei   BN   1 - 10   8. Bland Chromatin   BC   1
- 10   9. Normal Nucleoli   NN   1 - 10   10. Mitoses   Mit   1 -
10   11. Class: Class   (0 for benign, 1 for malignant)

k = 0.01

full outcome = 122.8882

deleted Size

new outcome = 120.8891

deleted SECS

new outcome = 119.2668

The variables used in this model are: Thick Shape Adh BN BC NN Mit

k = 0.05

full outcome = 122.8882

deleted Size

new outcome = 120.8891

deleted SECS

new outcome = 119.2668

deleted NN

new outcome = 121.7218

The variables used in this model are: Thick Shape Adh BN BC Mit

significance test approach

k = 0.01 or k = 0.05 (same)

Thick Adh BN BC

2c – Discrete Case n>5000 p<10 0-1Y Blocks Classification

use 1~10 attributes to predict the 11th attribute: class

https://archive.ics.uci.edu/ml/machine-learning-databases/page-blocks/

Number of Attributes height: integer. | Height of the block. lenght: integer. | Length of the block. area: integer. | Area of the block (height * lenght); eccen: continuous. | Eccentricity of the block (lenght / height); p_black: continuous. | Percentage of black pixels within the block (blackpix / area); p_and: continuous. | Percentage of black pixels after the application of the Run Length Smoothing Algorithm (RLSA) (blackand / area); mean_tr: continuous. | Mean number of white-black transitions (blackpix / wb_trans); blackpix: integer. | Total number of black pixels in the original bitmap of the block. blackand: integer. | Total number of black pixels in the bitmap of the block after the RLSA. wb_trans: integer. | Number of white-black transitions in the original bitmap of the block. **k = 0.01**
full outcome = 1636.061

deleted area new outcome = 1651.106 deleted mean_tr new outcome = 1653.132

The variables used in this model are: height lenght eccen p_black p_and blackpix blackand wb_trans

k = 0.05

deleted area new outcome = 1651.106 deleted mean_tr new outcome = 1653.132

deleted blackand new outcome = 1707.096 deleted blackpix new outcome = 1705.208

deleted wb_trans new outcome = 1708.491

The variables used in this model are: height lenght eccen p_black p_and

significance test approach

k = 0.01 or k = 0.05 (same)

all variables except for mean_tr

2c – Discrete Case n<1000 p>10 0-1 Y Wine Recognition Data

Use 2~14 attributes to predict 1st attribute: class
https://archive.ics.uci.edu/ml/datasets/Wine
k = 0.01

full outcome = 28 deleted Proline new outcome = 26 deleted Magnesium new outcome = 24 deleted intensity new outcome = 22 deleted phenols new outcome = 20 deleted Malic new outcome = 18

The variables used in this model are: Alcohol Ash Alcalinity Flavanoids Nonflavanoid Proanthocyanins Hue diluted

k = 0.05

full outcome = 28 deleted Proline new outcome = 26 deleted Magnesium
new outcome = 24deleted intensity new outcome = 22deleted phenols
new outcome = 20deleted Malic new outcome = 18

The variables used in this model are: Alcohol Ash Alcalinity Flavanoids Nonflavanoid Proanthocyanins Hue diluted

significance test approach
no variables

2c – n<1000 p<10 continuous Y

https://archive.ics.uci.edu/ml/datasets/Energy+efficiency
use 1~8 attributes to predict the 9th
k = 0.01
full outcome = 0.9154303

The variables used in this model are: X1 X2 X3 X4 X5 X6 X7 X8
k = 0.05
full outcome = 0.9154303
deleted X5
new outcome = 0.8978163
The variables used in this model are: X1 X2 X3 X4 X6 X7

ÿ significance test approach
k = 0.01 or k = 0.05 (same)
X1 X2 X3 X5 X7 X8

2c – n>5000 p<10 continuous Y Blocks Classification

use 2~9 attributes to predict the 1st attributek = 0.01 full outcome = 0.2802776
deleted F5
new outcome = 0.2800135
deleted F7
new outcome = 0.2789552
The variables used in this model are: F1 F2 F3 F4 F6 F8
k = 0.05
full outcome = 0.2802776
deleted F5
new outcome = 0.2800135
deleted F7
new outcome = 0.2789552
deleted F1
new outcome = 0.2703461
The variables used in this model are: F2 F3 F4 F6 F8
significance test approach
k = 0.01 or k = 0.05 (same)
F1 F2 F3 F4 F5 F6 F7 F8

2d – Leave one out strategy of PAC

Cross Validation – Method used to validate the effectiveness of different models by training the algorithm on a subset of data

Function – leave1out01() calculates the PAC value using the leave one out method

Function – predVal(x,y,predictors) evaluates the predicted value of Y based on logistic regression with 0.5 as the criterion

2d – Leave one out strategy of PAC

Using the “leaving one out “ to do the pima example, we get the same result
full outcome = 0.7682292
deleted Thick
new outcome = 0.7682292
deleted Insul
new outcome = 0.7695312
deleted BP
new outcome = 0.7695312
deleted Age
new outcome = 0.7708333
The variables used in this model are: NPreg Gluc BMI Genet