Bias and Parsimony in Regression Analysis

ECS 256 W14 Final Project Presentaion

Kevin Cosgrove, Wei Fang,
Xiaoyun Wang, Zhicheng Yang

Department of Computer Science
University of California, Davis

March 11, 2014

Outline

Problem Description

The population regression function is

$$m_{Y;X}(t) = t^{0.75} \quad t \in (0,1) \tag{1}$$

The estimated regression function is

$$\hat{m}_{Y;X}(t) = \beta t \quad t \in (0,1) \tag{2}$$

Find the asymptotic bias at t = 0.5.

SOLUTION

The key is Eqn.(23.34)

$$\hat{\beta} = (Q'Q)^{-1}Q'V$$

where in this case, $V = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$, $Q = (X_1, X_2, \cdots, X_n)$

plug into Eqn.(23.34),

$$\hat{\beta} = (\sum_{i=1}^{n} X_i^2)^{-1} \sum_{i=1}^{n} X_i Y_i \tag{3}$$

As the sample size n goes to infinity,

$$\beta = \frac{E(XY)}{E(X^2)} \tag{4}$$

SOLUTION (CONT.)

$$\beta = \frac{E(XY)}{E(X^2)}$$

The population regression function

$$m_{Y;X}(t) = t^{0.75} \quad t \in (0,1)$$

is equivalent to,

$$E(Y|X = t) = t^{0.75} \quad t \in (0,1) \tag{5}$$

$$E(Y|X) = X^{0.75} \quad X \sim U(0,1) \tag{6}$$

$$E(XY) = E[E(XY|X)] = E[XE(Y|X)] = E(X^{1.75})$$

$$E(X^{1.75}) = \int_0^1 t^{1.75} f_X(t)dt = \int_0^1 t^{1.75} dt = \frac{1}{2.75}$$

$$E(X^2) = \int_0^1 t^2 f_X(t)dt = \int_0^1 t^2 dt = \frac{1}{3}$$

SOLUTION (CONT.)

$$\beta = \frac{3}{2.75} = 1.090909091$$

The bias function is

$$\begin{aligned} bias(t) &= E[\hat{m}_{Y;X}(t)] - m_{Y;X}(t) && (7)\\ &= E(\beta t) - t^{0.75} && (8)\\ &= 0.5\beta - t^{0.75} \quad t \in (0,1) && (9) \end{aligned}$$

At $t = 0.5$ the bias is

$$bias(t = 0.5) = -0.04914901$$

Outline

Problem 1

Bias Of An Approximate Regression Model

Problem 2

a. Parsimony
b. Testing On Simulated Data
c. Testing On Real Data Sets
d. Another PAC Function

Problem 2a. Parsimony

- Goal: Develop a model selection method that yields parsimony no matter how large the sample data is.
- Function Declarations:

```
prsm(y,x,k=0.01,predacc=ar2,crit,printdel=F)
ar2(y,x)
aiclogit(y,x)
compare(y,x,predacc)
```

- In prsm(), predictor variables are deleted in the least "significant" order.
- ar2() is a "max" PAC function.
  - New PAC value is acceptable if $> (1 - k)$PAC.
- aiclogit() is a "min" PAC function.
  - New PAC value is acceptable if $< (1 + k)$PAC.

Problem 2b. Testing On Simulated Data

Table : Recommended Predictor Set

Sample size	Runs	Parsimony Model		Significance Testing
		k=0.01	k=0.05	
100	1	1 2 3 9	1 2 3	1 2 3 9
	2	1 2 3 6 7 9	1 2 3 6 7 9	1 2 3 7
	3	1 2 3	1 2 3	1 2 3
1000	1	1 2 3	1 2 3	1 2 3 4
	2	1 2 3	1 2 3	1 2 3
	3	1 2 3	1 2 3	1 2 3
10000	1	1 2 3	1 2 3	1 2 3 4
	2	1 2 3	1 2 3	1 2 3 4 9
	3	1 2 3	1 2 3	1 2 3 4
100000	1	1 2 3	1 2 3	1 2 3 4 7
	2	1 2 3	1 2 3	1 2 3 4
	3	1 2 3	1 2 3	1 2 3 4 8

Problem 2c. Testing On Real Data Sets

Data set criteria:

- Small n (< 1000), small p (< 10), continuous Y
  - Data Set #1: Concrete Compressive Strength
- Small n (< 1000), small p (< 10), 0-1 Y
  - Data Set #2: Pima Indians Diabetes
- Small n (< 1000), large p (> 15), continuous Y
  - Data Set #3: Parkinsons
- Small n (< 1000), large p (> 15), 0-1 Y
  - Data Set #4: Ionosphere
- Large n (> 5000), small p (< 10), continuous Y
  - Data Set #5: Wine Quality
- Large n (> 5000), small p (< 10), 0-1 Y
  - Data Set #6: Page Blocks Classification
- Large n (> 5000), large p (> 15), continuous Y
  - Data Set #7: Waveform Database Generator
- Large n (> 5000), large p (> 15), 0-1 Y
  - Data Set #8: EEG Eye State

Data Set #1: Concrete Compressive Strength

- Small n = 1030, small p = 9, continuous Y
- This data set consists of 7 concrete mixtures' component densities, the age since it was poured, and its compressive strength. The densities and the age are the set's predictor variables (total of 8), and the strength is the response variable.
- We chose to use the ar2 PAC function with $k = 0.01$ and 0.05, as well as significance testing with $\alpha = 5\%$. These tests deleted 3, 3, and 2 predictor variables, respectively.

Table : Test Result On Data Set # 1

Date Set #	Parsimony Model		Significance Testing
	k=0.01	k=0.05	
1	1 2 3 4 8	1 2 3 4 8	1 2 3 4 5 8

Data Set #2: Pima Indians Diabetes

- Small n = 768, small p = 8, 0-1 Y
- This data set consists of 8 different medical measures of Pima Indian women over the age of 21, and a boolean class variable.
- We chose to use the AIC PAC function with k = 0.01 and 0.05, and significance testing with α = 5%. These tests deleted 4, 7, and 3 predictor variables, respectively.

Table : Test Result On Data Set # 2

Date Set #	Parsimony Model		Significance Testing
	k=0.01	k=0.05	
2	1 2 6 7	2	1 2 3 6 7

Data Set #3: Parkinsons

- Small n = 197, large p = 23, continuous Y
- This data set is composed of 22 medical measures of patients with or without Parkinson's disease. The predictor varaibles are the results of the medical tests and the response variable is a boolean for the presence of Parkinson's.
- We chose to use the ar2 PAC function with k = 0.01 and 0.05, and significance testing with α = 5%. These tests deleted 11, 15, and 19 predictor variables, respectively.

Table : Test Result On Data Set # 3

Date Set #	Parsimony Model		Significance Testing
	k=0.01	k=0.05	
3	1 3 4 8 9 12 15 16 17 19 20	1 4 8 19 20	4 17 20

Data Set #4: Ionosphere

- Small n = 351, large p = 34, 0-1 Y
- This data set consists of measurements of electromagnetic tests in the ionosphere and a boolean class value.
- The second column for the data set was all zeros.
- We chose to use the AIC PAC function with k = 0.01 and 0.05, and significance testing with α = 5%. These tests deleted 15, 24, and 20 predictor variables, respectively.

Table : Test Result On Data Set # 4

Date Set #	Parsimony Model		Significance Testing
	k=0.01	k=0.05	
4	1 4 5 7 8 10 14 15 17 18 21 22 24 26 28 29 30 33	1 4 5 7 14 21 26 28 29 33	1 2 4 6 7 8 18 21 22 25 26 30 33

Data Set #5: Wine Quality

- Large n = 4898, small p = 12, continuous Y
- This data set is composed of measures of different types of white wine. The response variable is a rating tasting score between 0 and 10, and the 11 predictor variables are various chemical measures.
- We chose to use the ar2 PAC function with k = 0.01 and 0.05, and significance testing with α = 5%. These tests deleted 4, 8, and 3 predictor variables, respectively.

Table : Test Result On Data Set # 5

Date Set #	Parsimony Model		Significance Testing
	k=0.01	k=0.05	
5	1 3 4 8	1 3 4	1 2 3 4 5 6 7 8 9

Data Set #6: Page Blocks Classification

- Large n = 5473, small p = 10, 0-1 Y
- This data set consists of 11 different measures relating to the amount of black and white space in parts of different text documents. None of the variables are inherently response variables, but we chose the number of white-black transitions to be the response variable for our tests.
- We chose to use the AIC PAC function with k = 0.01 and 0.05, and significance testing with α = 5%. These tests deleted 3, 5, and zero predictor variables, respectively.

Table : Test Result On Data Set # 6

Date Set #	Parsimony Model		Significance Testing
	k=0.01	k=0.05	
6	1 2 3 4 5 6 10	1 2 4 5 6	1 2 3 4 5 6 7 8 9 10

Data Set #7: Waveform Database Generator

- Large n = 5000, large p = 40, continuous Y
- This data set is composed of 40 predictor variables which are different measures of waves, about half of which are normalized. The response variable is one of 3 different types of waves.
- We chose to use the ar2 PAC function with k = 0.01 and 0.05, and significance testing with α = 5%. These tests deleted 34, 37, and 25 predictor variables, respectively.

Table : Test Result On Data Set # 7

Date Set #	Parsimony Model		Significance Testing
	k=0.01	k=0.05	
7	5 6 10 11 12 13	16 11 12	3 4 5 6 7 9 10 11 12 13 14 15 17 18 19

Data Set #8: EEG Eye State

- Large n = 14980, large p = 15, 0-1 Y
- This data set consists of 14 measures of an EEG test with the response variable a boolean indicating whether the subject's eyes were open or closed.
- We chose to use the AIC PAC function with k = 0.01 and 0.05, and significance testing with α = 5%. These tests deleted 10, 13, and 1 variables, respectively.

Table : Test Result On Data Set # 8

Date Set #	Parsimony Model		Significance Testing
	k=0.01	k=0.05	
8	1 2 5 6	2	1 2 3 4 5 6 7 9 10 11 12 13 14

Problem 2d. Another PAC Function

- Leave-one-out cross-validation.
- PAC value is the proportion of correct classfications. So this is a "max" PAC function.
- The PAC function's running time is linear with the sample size.
- Two implementations:
  - Self-made cross-validation: For each observation in the sample data, we temporarily delete it from training set, and reserve it as the validation set. Perform the training-validation process though every observation, count the number of correct classifications. Return the proportion of correct predictions.
  - Use R's cv.glm() function in boot package.

Problem 2d. Another PAC Function (Cont.)

Output 1:

```
full outcome = 0.7682292
deleted V4
new outcome = 0.7682292
deleted V5
new outcome = 0.7695312
deleted V8
new outcome = 0.7721354
deleted V3
new outcome = 0.7708333
[1] 1 2 6 7
```

Output 2:

```
full outcome = 0.7773437
deleted V4
new outcome = 0.7773437
deleted V1
new outcome = 0.7747396
deleted V5
new outcome = 0.7734375
deleted V3
new outcome = 0.7734375
deleted V8
new outcome = 0.7695312
deleted V7
new outcome = 0.7630208
[1] 2 6
```

References

- UCI Machine Learning Repository: Concrete Compressive Strength Data Set
  https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength
- UCI Machine Learning Repository: Pima Indians Diabetes Data Set
  https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
- UCI Machine Learning Repository: Parkinsons Data Set
  https://archive.ics.uci.edu/ml/datasets/Parkinsons
- UCI Machine Learning Repository: Ionosphere Data Set
  https://archive.ics.uci.edu/ml/datasets/Ionosphere
- UCI Machine Learning Repository: Wine Quality Data Set
  https://archive.ics.uci.edu/ml/datasets/Wine+Quality

References (cont.)

- UCI Machine Learning Repository: Page Blocks Classification Data Set https://archive.ics.uci.edu/ml/datasets/Page+Blocks+Classification
- UCI Machine Learning Repository: Waveform Database Generator (Version 2) Data Set https://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+%28Version+2%29
- UCI Machine Learning Repository: EEG Eye State Data Set https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State