Correct Model Selection in Multiple Regression Analyses of Big Data

Katherine Thompson

Department of Statistics, University of Kentucky

October 17, 2017

MOTIVATION

Goals:

- Improve statistical modeling in a variety of application areas
- Correctly identify the relationships present in data sets
- Understand the difficulty in choosing the correct statistical model in big data

OUTLINE

- Introduction to the Challenges
- Methods
- Results
- Conclusions and Future Directions

MOTIVATING EXAMPLE: NHANES Data (Hofe et al. 2014)

Goal: Identify variables related to HDL cholesterol

Data Set:

- Sample Size: $n = 5038$
- Variables: 176

Challenges:

- Big data

MOTIVATING EXAMPLE: NHANES Data (Hofe et al. 2014)

Goal: Identify variables related to HDL cholesterol

Data Set:

- Sample Size: $n = 5038$
- Variables: 176

Challenges:

- Big data
- Small effects

MOTIVATING EXAMPLE: NHANES Data (Hofe et al. 2014)

Goal: Identify variables related to HDL cholesterol

Data Set:

- Sample Size: $n = 5038$
- Variables: 176

Challenges:

- Big data
- Small effects
- Complicated relationships

MOTIVATING EXAMPLE: NHANES Data (Hofe et al. 2014)

Subject	Diabetes	LBX118LA	BMXBMI	Serum Carotenoids	RIDRETH1
1	0	17.17	31.26	2.29	three
2	0	7.50	25.49	1.34	three
3	0	8.50	19.60	1.48	four
4	0		28.32	0.93	three
5	0	3.20	19.34	1.90	one
6	0		16.57		four
7	0	3.00	38.03	1.12	one
8	0	12.70	22.55	1.39	four

METHODS FOR MODEL SELECTION

Existing Methods Include:

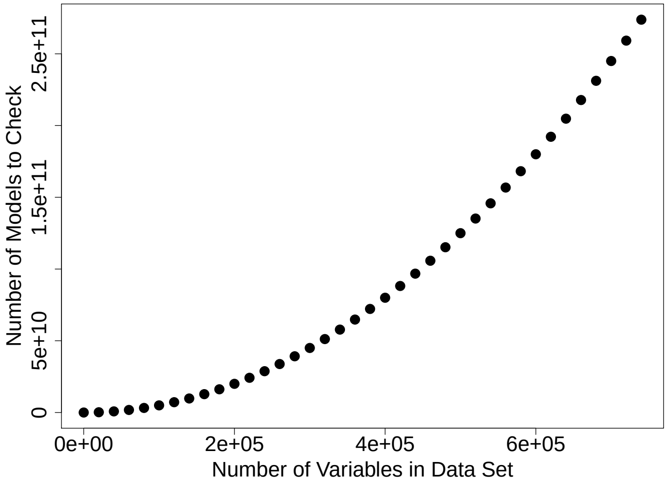
- Forward and Backward Selection

METHODS FOR MODEL SELECTION

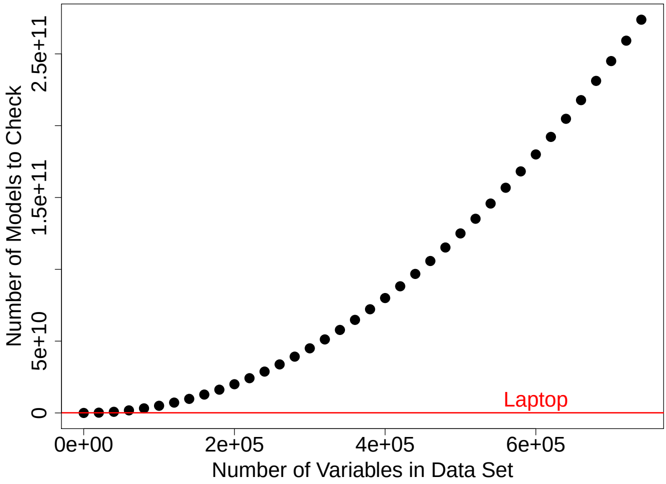
Existing Methods Include:

- Forward and Backward Selection
- Subset Selection or Exhaustive Search Methods

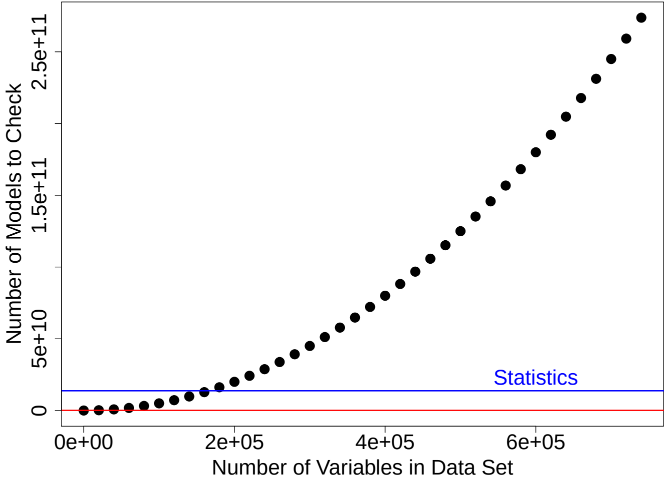
METHODS: Exhaustive Search



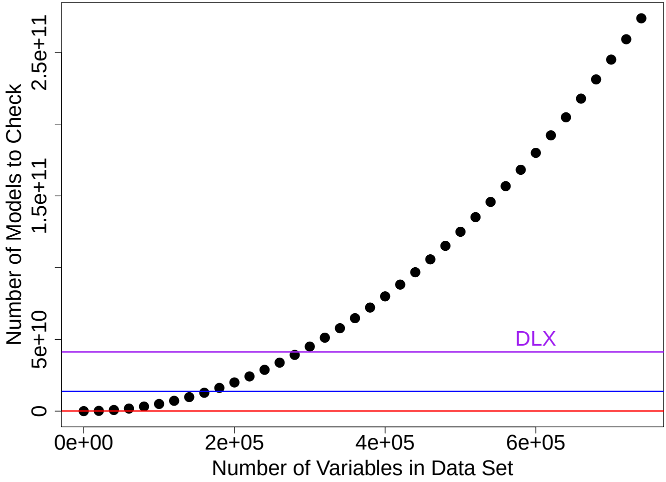
METHODS: Exhaustive Search



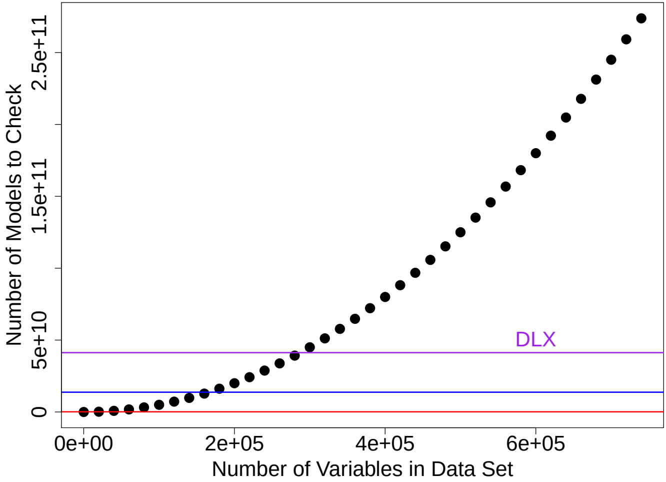
METHODS: Exhaustive Search



METHODS: Exhaustive Search



METHODS: Exhaustive Search



Even if we can check all possible models, what are the chances of identifying the correct model?

METHODS

Goal: Calculate this probability as a function of the effects and variation in the data.

METHODS

Goal: Calculate this probability as a function of the effects and variation in the data.

- Method: Multiple Linear Regression to Identify Optimal Model

METHODS

Goal: Calculate this probability as a function of the effects and variation in the data.

- Method: Multiple Linear Regression to Identify Optimal Model
- Compare to an alternative method

METHODS

Goal: Calculate this probability as a function of the effects and variation in the data.

- Method: Multiple Linear Regression to Identify Optimal Model
- Compare to an alternative method
- Computation: Use University of Kentucky High Performance Computing Center supercomputer

ALTERNATIVE METHOD: Feasible Solutions Algorithm

Feasible Solutions Algorithm (FSA): (Lambert 2016)

- Fast, flexible search algorithm

ALTERNATIVE METHOD: Feasible Solutions Algorithm

Feasible Solutions Algorithm (FSA): (Lambert 2016)

- Fast, flexible search algorithm
- Stochastic in starting point

ALTERNATIVE METHOD: Feasible Solutions Algorithm

Feasible Solutions Algorithm (FSA): (Lambert 2016)

- Fast, flexible search algorithm
- Stochastic in starting point
- Can produce multiple possible models for further exploration

ALTERNATIVE METHOD: FSA Example

Challenge: Suppose we are interested in modeling diabetes risk using two of the following: height, weight, age, sex, and diet

1. Randomly select two variables and fit the following model:
    - Diabetes $\sim$ height and age

ALTERNATIVE METHOD: FSA Example

Challenge: Suppose we are interested in modeling diabetes risk using two of the following: height, weight, age, sex, and diet

1. Randomly select two variables and fit the following model:
    - Diabetes $\sim$ height and age
2. "Swap" variables to find a better model

ALTERNATIVE METHOD: FSA Example

Challenge: Suppose we are interested in modeling diabetes risk using two of the following: height, weight, age, sex, and diet

1. Randomly select two variables and fit the following model:
   - Diabetes $\sim$ height and age
2. "Swap" variables to find a better model
   - Diabetes $\sim$ height and age
   - Diabetes $\sim$ weight and age

ALTERNATIVE METHOD: FSA Example

Challenge: Suppose we are interested in modeling diabetes risk using two of the following: height, weight, age, sex, and diet

1. Randomly select two variables and fit the following model:
   - Diabetes $\sim$ height and age
2. "Swap" variables to find a better model
   - Diabetes $\sim$ height and age
   - Diabetes $\sim$ weight and age
   - Diabetes $\sim$ sex and age

ALTERNATIVE METHOD: FSA Example

Challenge: Suppose we are interested in modeling diabetes risk using two of the following: height, weight, age, sex, and diet

1. Randomly select two variables and fit the following model:
   - Diabetes $\sim$ height and age
2. "Swap" variables to find a better model
   - Diabetes $\sim$ height and age
   - Diabetes $\sim$ weight and age
   - Diabetes $\sim$ sex and age
   - Diabetes $\sim$ diet and age

ALTERNATIVE METHOD: FSA Example

Challenge: Suppose we are interested in modeling diabetes risk using two of the following: height, weight, age, sex, and diet

1. Randomly select two variables and fit the following model:
   - Diabetes $\sim$ height and age
2. "Swap" variables to find the best model
   - Diabetes $\sim$ sex and age
3. Swap the remaining variable in the model.

ALTERNATIVE METHOD: FSA Example

Challenge: Suppose we are interested in modeling diabetes risk using two of the following: height, weight, age, sex, and diet

1. Randomly select two variables and fit the following model:
   - Diabetes $\sim$ height and age
2. "Swap" variables to find the best model
   - Diabetes $\sim$ sex and age
3. Swap the remaining variable in the model.
   - Diabetes $\sim$ sex and diet

ALTERNATIVE METHOD: FSA Example

Challenge: Suppose we are interested in modeling diabetes risk using two of the following: height, weight, age, sex, and diet

1. Randomly select two variables and fit the following model:
   - Diabetes $\sim$ height and age
2. "Swap" variables to find the best model
   - Diabetes $\sim$ sex and age
3. Swap the remaining variable in the model.
   - Diabetes $\sim$ sex and diet
4. Continue this process until we can not improve the model, resulting in a possible model.
   - **Diabetes $\sim$ sex and diet**

RESULTS: Simulated Linear Regression Data

For each simulated data set:

- Analyze by calculating the probability that the underlying correct model is the optimal model

RESULTS: Simulated Linear Regression Data

For each simulated data set:

- Analyze by calculating the probability that the underlying correct model is the optimal model
- Analyze using FSA and record if any feasible solution is the correct model (underlying truth)

RESULTS: Simulated Linear Regression Data

For each simulated data set:

- Analyze by calculating the probability that the underlying correct model is the optimal model
- Analyze using FSA and record if any feasible solution is the correct model (underlying truth)

Notation:

- $\sigma^2 =$ variance of error terms in the regression model

RESULTS: Simulated Linear Regression Data

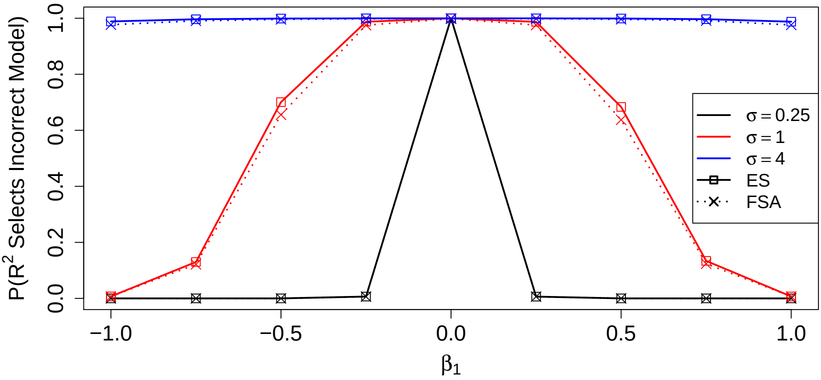
For each simulated data set:

- Analyze by calculating the probability that the underlying correct model is the optimal model
- Analyze using FSA and record if any feasible solution is the correct model (underlying truth)

Notation:

- σ^2 = variance of error terms in the regression model
- β_1 = coefficient values for each regression data set

RESULTS: Simulated Linear Regression Data



RESULTS: Example Based on NHANES Data

Data for LBXHDD:

- Removed existing detectable effects from the data

RESULTS: Example Based on NHANES Data

Data for LBXHDD:

- Removed existing detectable effects from the data

Created True Model for LBXHDD:

- Hide effects of LBXD01LA, DR2TVK, LBXD01LA*DR2TVK on LBXHDD in data
- Small effects of each covariate

RESULTS: Example Based on NHANES Data

Data for LBXHDD:

- Removed existing detectable effects from the data

Created True Model for LBXHDD:

- Hide effects of LBXD01LA, DR2TVK, LBXD01LA*DR2TVK on LBXHDD in data
- Small effects of each covariate

Model Identified by Exhaustive Search:

- RIDRETH1.3, status2, RIDRETH1.3*status2

RESULTS: Example Based on NHANES Data

Data for LBXHDD:

- Removed existing detectable effects from the data

Created True Model for LBXHDD:

- Hide effects of LBXD01LA, DR2TVK, LBXD01LA*DR2TVK on LBXHDD in data
- Small effects of each covariate

Model Identified by Exhaustive Search:

- RIDRETH1.3, status2, RIDRETH1.3*status2

Models Identified by FSA:

- RIDRETH1.3, status2, RIDRETH1.3*status2
- LBXD01LA, DR2TVK, LBXD01LA*DR2TVK

CONCLUSIONS AND FUTURE DIRECTIONS

Conclusions:

- Using the statistically optimal model results in the incorrect model selection a large percentage of the time.
- FSA can identify correct models in the potential variable sets, even in cases when exhaustive search procedures do not.

Future Directions:

- Consider analyzing models with more than two variables and/or higher order interactions.
- Derive a hypothesis to test that a selected model is correct.

References and Contact Information

Acknowledgements:

- Thanks to the University of Kentucky High Performance Computing Center for the use of the supercomputer for simulation data analysis.

References:

- Hofe, Carolyn R., et al. "Fruit and vegetable intake, as reflected by serum carotenoid concentrations, predicts reduced probability of polychlorinated biphenylassociated risk for type 2 diabetes: National Health and Nutrition Examination Survey 2003-2004." Nutrition research 34.4 (2014): 285-293.
- Joshua Lambert (2016). rFSA: Feasible Solution Algorithm for Finding Best Subsets and Interactions. R package version 0.1.1. `https://CRAN.R-project.org/package=rFSA`

Contact Information:

- Katherine Thompson: katherine.thompson@uky.edu