Simple Bootstrap and Simulation Approaches to Quantifying Reliability of High-Dimensional Feature Selection

Frank E Harrell Jr

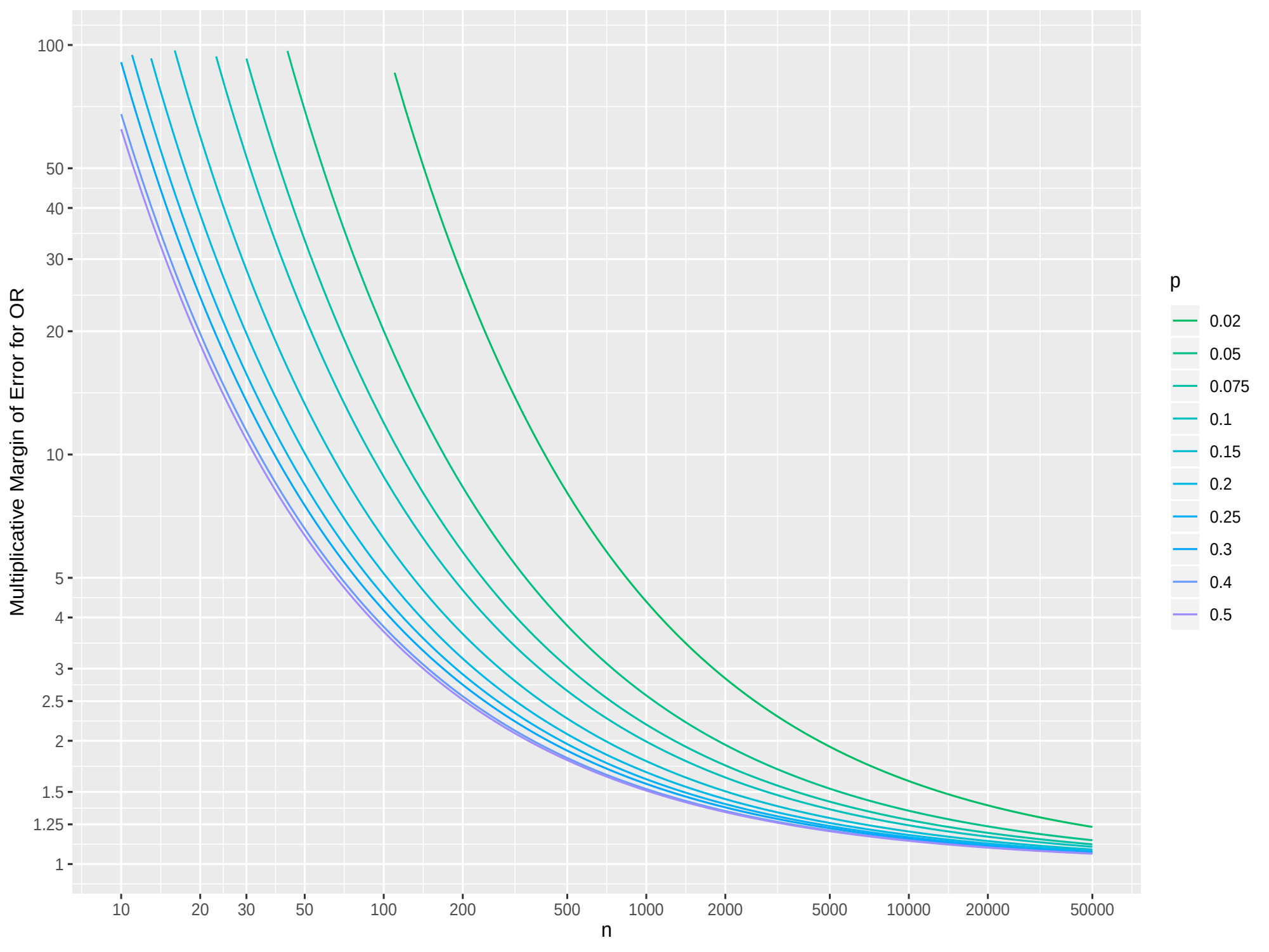
Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, Tennessee

Statistical Advisor, Office of Biostatistics
US FDA Center for Drug Evaluation and Research

JSM 2018 ENAR Session Vancouver BC 2018-07-31

Simplest Task: Estimate Association Between X_1 and Y With Precision

- Correlation coefficient, hazard ratio, odds ratio, etc.
- Consider multiplicative margin of error for an odds ratio
- Assume true $P(Y = 1)$ in each group be $\geq p$
- Assume total sample size n is split into $\frac{1}{10}$ of subjects with $X = 0$ and $\frac{9}{10}$ with $X = 1$
- MMOE = anti-log of half-width of 0.95 CL on log OR

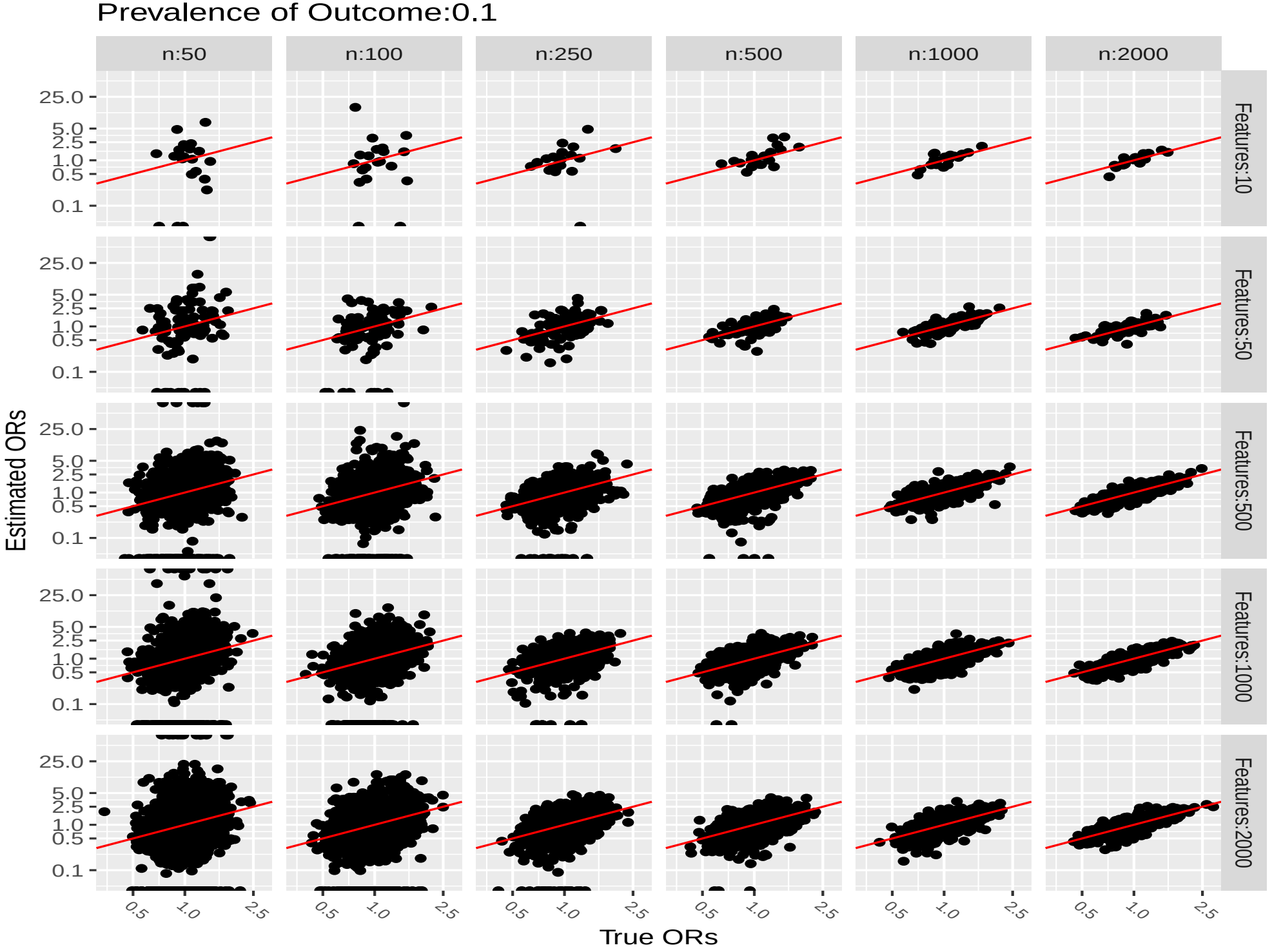

Multiplicative Margin of Error for OR
100
50
40
30
20
10
5
4
3
2.5
2
1.5
1.25
1
10
20
30
50
100
200
500
1000
2000
5000
10000
20000
50000
n
p
0.02
0.05
0.075
0.1
0.15
0.2
0.25
0.3
0.4
0.5

The Real Task: High-Dimensional Modeling or Feature Selection

- One-at-a-time feature selection from among large # candidate features
- Simultaneous modeling of large # features
- Ranking importance of features
- False discovery risks ignore false negative risks and precision of final estimates
- Use simple simulation to understand needed sample sizes

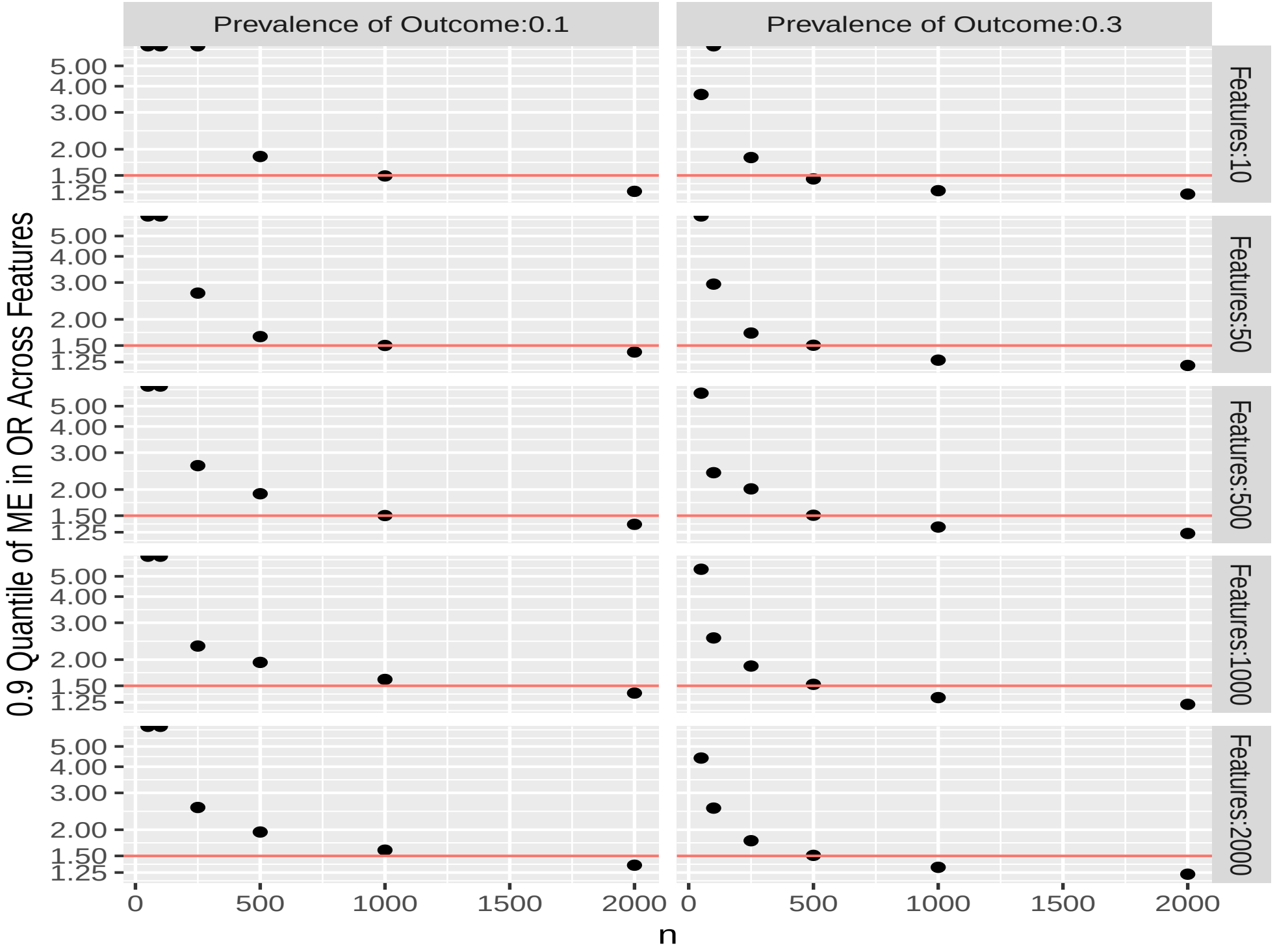
Controlling Margins of Error of Entire Set of OR Estimates

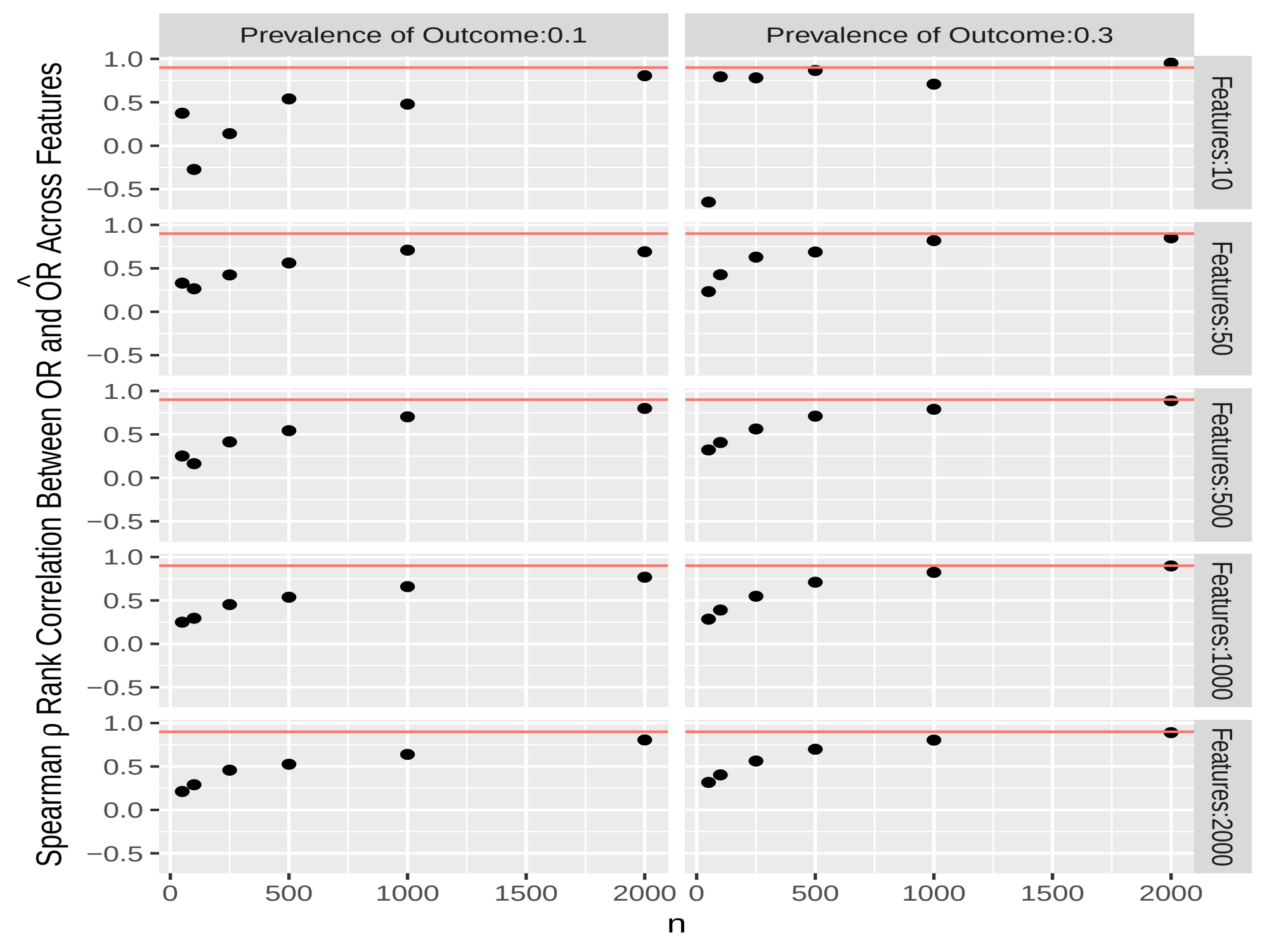
- Let $p = \#$ of candidate binary features
- Assume true log ORs have normal($\mu = 0, \sigma = 0.25$) distribution
- Want to
  - judge ability to jointly estimate p associations
  - rank order features by observed associations
- Y prevalence 0.1, X_i prev. $\sim U(0.05, 0.5)$


Prevalence of Outcome:0.1
n:50
n:100
n:250
n:500
n:1000
n:2000
Features:10
Features:50
Features:500
Features:1000
Features:2000
Estimated ORs
True ORs
25.0
5.0
2.5
1.0
0.5
0.1
0.5
1.0
2.5

Summarizing Multiplicative Errors in ORs

- ME = anti-log of estimate / true OR
- Compute 0.9 quantile of ME over whole set of estimated ORs
- Then compute Spearman rank correlation between estimated and true underlying ORs over whole set


Prevalence of Outcome:0.1
Prevalence of Outcome:0.3
Features:10
Features:50
Features:500
Features:1000
Features:2000
0.9 Quantile of ME in OR Across Features
5.00
4.00
3.00
2.00
1.50
1.25
0
500
1000
1500
2000
n

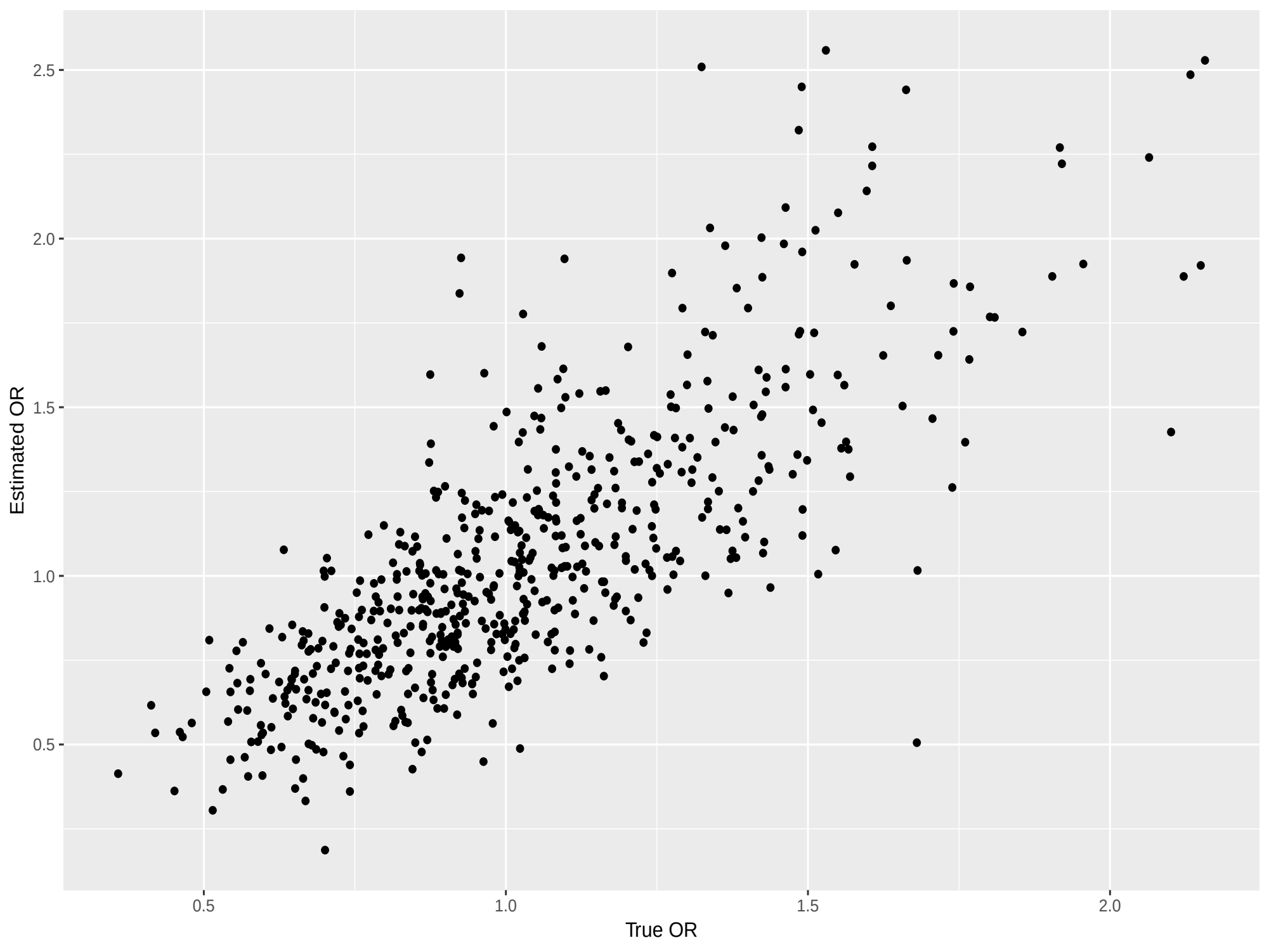

Prevalence of Outcome:0.1
Prevalence of Outcome:0.3
Features:10
Features:50
Features:500
Features:1000
Features:2000
Spearman ρ Rank Correlation Between OR and ÔR Across Features
n
1.0
0.5
0.0
−0.5
0
500
1000
1500
2000

Bootstrap Analysis for One Simulated Dataset

- Are winning features really winners? Are losers really losers?
  - Compute confidence intervals for feature importance ranking
- What is the bias in an OR that passes the selection process?
- Bootstrap can take into account all sources of uncertainty
  - Actually underestimates instability compared to exact Bayesian credible intervals

Bootstrap Analysis, *continued*

- Simulate data for 600 subjects with 600 candidate predictors
- Estimate bias in apparent lowest and apparent highest ORs
  - Find extreme estimated ORs in sample w/replacement
  - Compute dropoff when compared to same predictor in orig. sample
- Y has prevalence 0.2


Estimated OR
2.5
2.0
1.5
1.0
0.5
0.5
1.0
1.5
2.0
True OR

Features with Largest and Smallest Estimated ORs

Largest observed OR

Estimated OR: 2.56, feature #1
true OR: 1.53, ME: $\times 1.67$

Smallest observed OR

Estimated OR: 0.19, feature #355
true OR: 0.7, ME: $\times 0.27$

Bootstrap Bias Estimates

Largest OR

0.95 CL for rank: 440 600
Median bias: 1.64 Geometric mean bias: 1.69
True ME: 1.67

Bootstrap bias-corrected OR: 1.56
Original OR: 2.56 True OR: 1.53

Smallest OR

0.95 CL for rank: 1 45
Median bias: 0.24 Geometric mean bias: 0.11
True ME: 0.27

Bootstrap bias-corrected OR: 0.76
Original OR: 0.19 True OR: 0.7

Summary

- Simple simulations and bootstrap can determine limits in information content from a sample of low or high-dimensional data
- Recognize that feature selection (not always recommended) is a ranking problem
- Bootstrap CLs for predictor ranks provide basis or caution in feature selection
- Bootstrap can estimate bias when selecting on extreme observed associations
- One can estimate the sample size so that data are unlikely to be tortured beyond the breaking point