EXAMINING THE PERFORMANCE OF REGULARIZATION PENALTIES FOR VARIABLE SELECTION IN THE BETA REGRESSION MODEL

Alfonso J. Martinez (PhD student, EMS; MS student, Statistics)

GSEC Graduate Research Colloquium

10/28/22

DO YOU COLLECT DATA OR ANALYZE DATA WITH QUANTITATIVE METHODS (E.G., REGRESSION)?

DO YOU KNOW WHICH VARIABLES ARE IMPORTANT/ASSOCIATED WITH AN OUTCOME VARIABLE PRIOR TO ANALYSIS?

WHICH VARIABLES ARE IMPORTANT FOR EXPLAINING STUDENT TRUANCY?

Outcome (dependent) variable

Student truancy (proportion)

Candidate covariates

- Parental involvement
- Height
- Free and reduced lunch
- Learning disability
- Number of friends
- Age

MODEL SELECTION

- Process of selecting the best model from all the available models for a particular problem based on different criterions such as robustness and model complexity.

VARIABLE SELECTION

- Process of selecting the best subset of predictors/covariates for a given problem
- Different from model selection: select one specific model from the list of available predictive models
- Failure to identify covariates that are related to the outcome leads to incorrect inferences
- Retaining redundant/insignificant covariates decreases a model's predictive accuracy and makes model unnecessarily complex

REGULARIZATION VIA PENALTY FUNCTIONS

- Continuous variable selection via penalty functions
- Penalty functions penalize unimportant covariates in the model
- Dual purpose:
  - Remove unimportant covariates from the model WHILE
  - Simultaneously estimating the coefficients of the important covariates
- Oracle property: perform as well as if the analyst had known in advance which covariates were not important and which were important

PENALTY FUNCTIONS

Smoothly–clipped absolute deviation (SCAD)

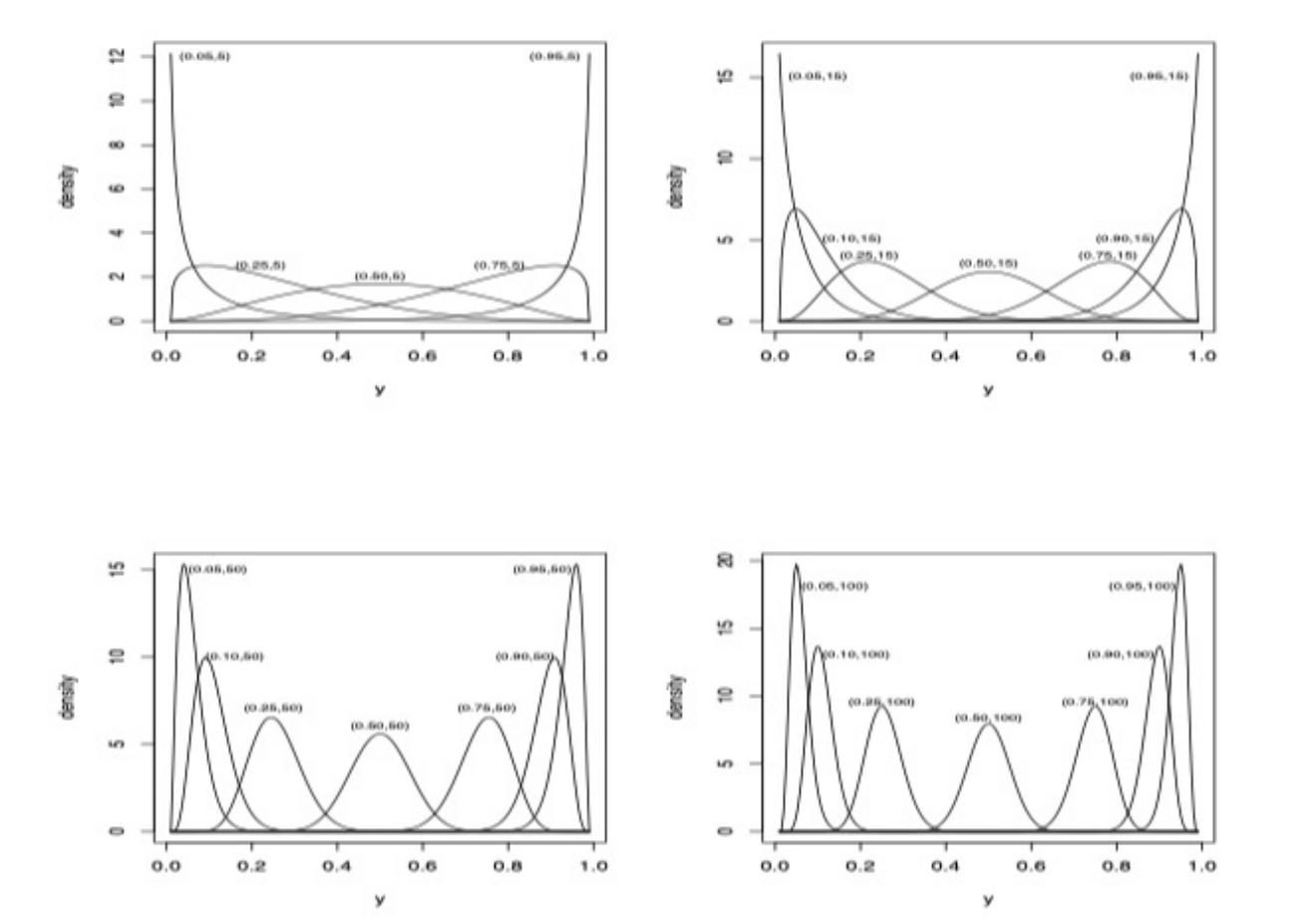
$$P(x|\lambda,\gamma) = \begin{cases} \lambda|x| & \text{if } |x| \le \lambda, \\ \frac{2\gamma\lambda|x| - x^2 - \lambda^2}{2(\gamma-1)} & \text{if } \lambda < |x| < \gamma\lambda, \\ \frac{\lambda^2(\gamma+1)}{2} & \text{if } |x| \ge \gamma\lambda \end{cases}$$

Minimax Concave Penalty (MCP)

$$P_\gamma(x;\lambda) = \begin{cases} \lambda|x| - \frac{x^2}{2\gamma}, & \text{if } |x| \le \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & \text{if } |x| > \gamma\lambda \end{cases}$$

CASE STUDY: BETA REGRESSION

- Beta regression model: generalized linear model for modeling rates/proportions (outcome is value between 0 and 1, i.e., truancy).
- Popular in many social science disciplines, including psychology and education



RESEARCH QUESTION

- How well do the SCAD and MCP penalty functions perform in the beta regression model when there are varying levels of multicollinearity in the covariates?
- Multicollinearity = correlation among the covariates

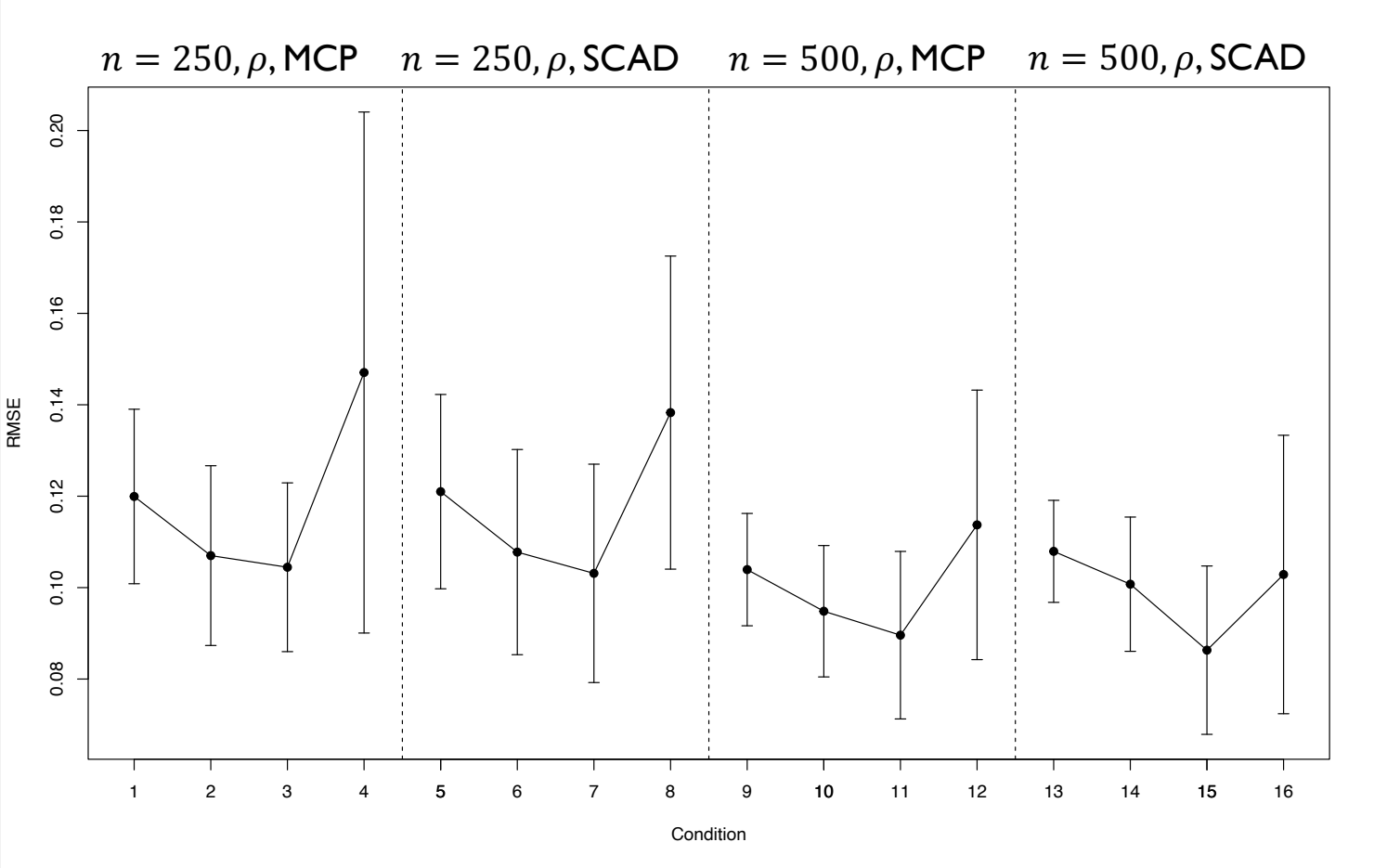
RATIONALE

- Beta regression is becoming increasingly popular; however, no studies have examined variable selection in this setting
- It is unknown how correlation among covariates influences the SCAD and MCP penalty functions, if at all

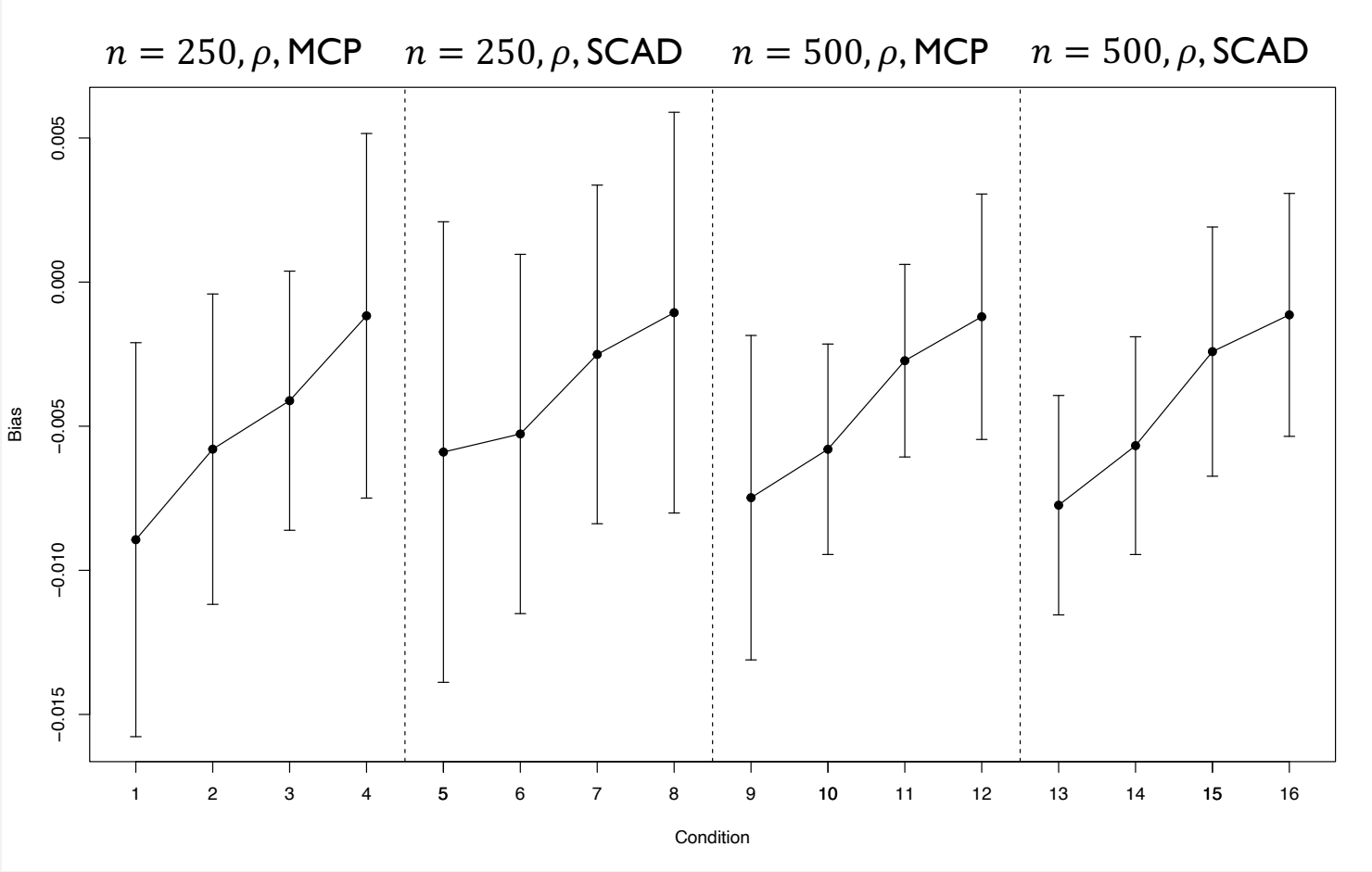
SIMULATION STUDY

- Fully-crossed design with three manipulated conditions:
  - Penalty (SCAD, MCP)
  - Sample size (250, 500)
  - Correlation between covariates (0, 0.3, 0.6, 0.9)
- Data generated with $p = 30$ covariates
- True coefficients: $\beta = (-2, -1.5, -1, -0.5, 0.5, 1, 1.5, 2, 0_{21\times 1})$
- 100 replications per condition
- Outcomes I: root mean square error, bias, BIC
- Outcomes II: average number of zero coefficients that were correctly dropped

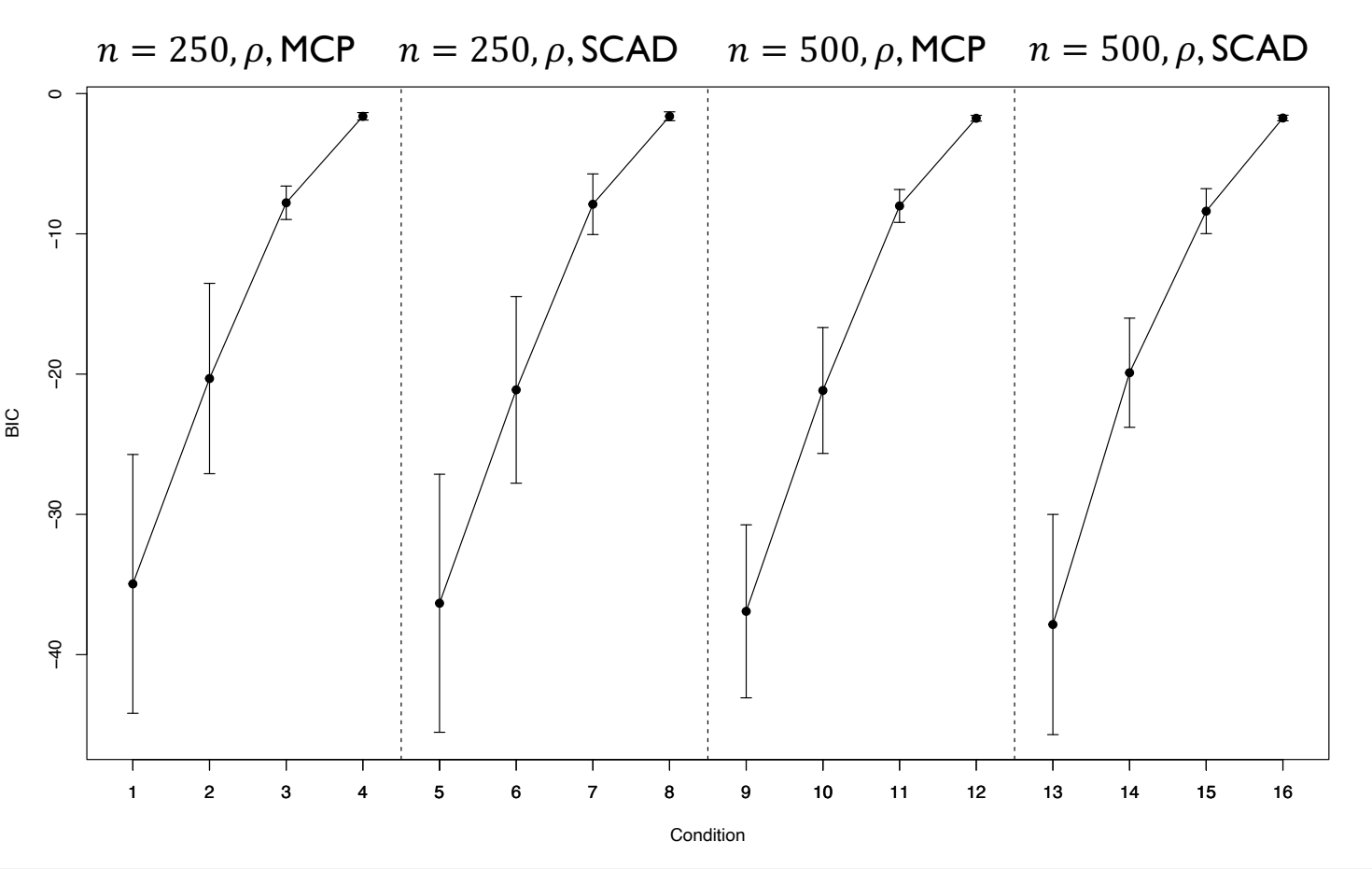
RESULTS: RMSE



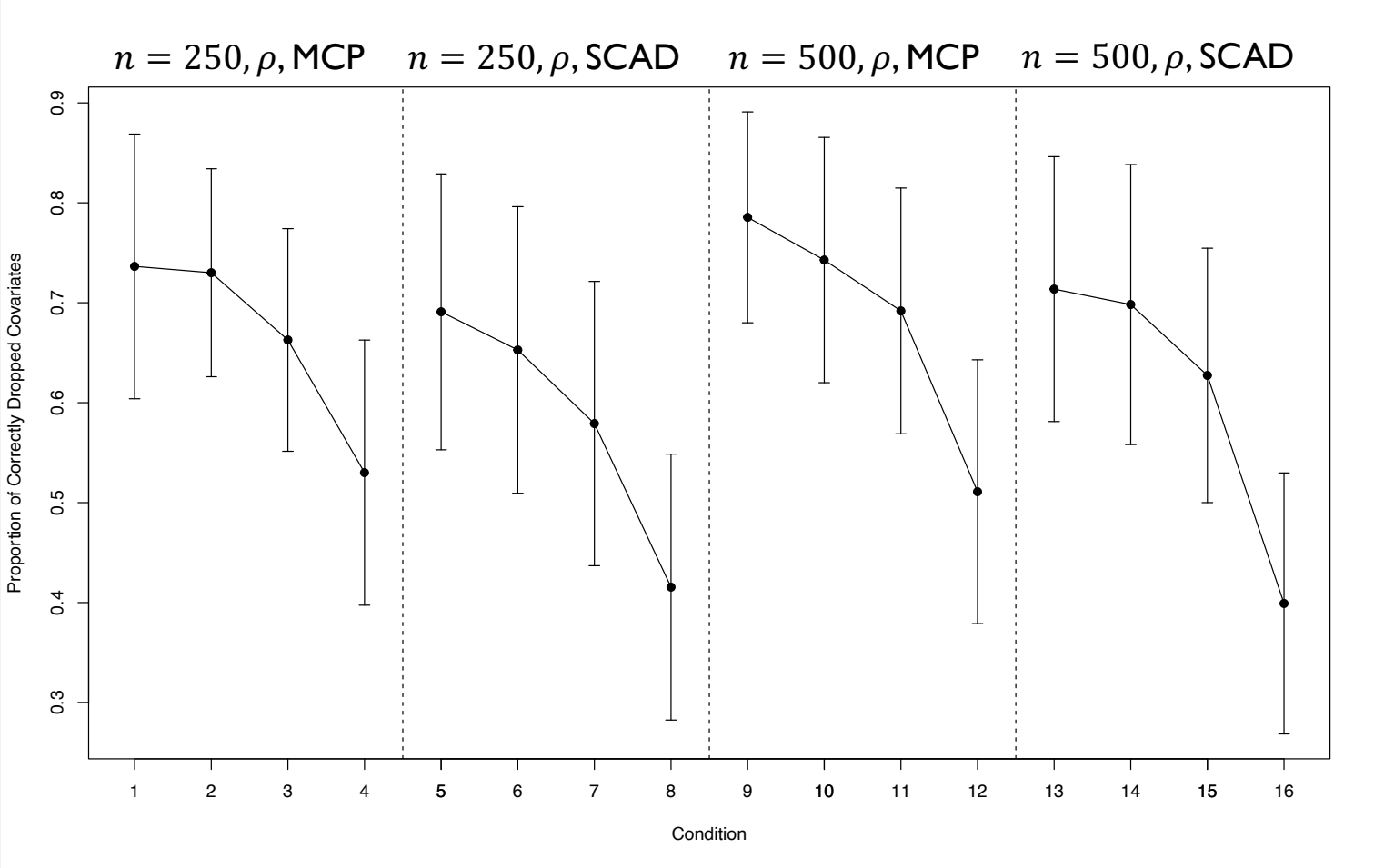
RESULTS: BIAS



RESULTS: BIC


$n = 250, \rho, \text{MCP}$
$n = 250, \rho, \text{SCAD}$
$n = 500, \rho, \text{MCP}$
$n = 500, \rho, \text{SCAD}$
BIC
0
−10
−20
−30
−40
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Condition

RESULTS: PROPORTION OF CORRECTLY DROPPED COVARIATES



CONCLUSION

- RMSE decreased as sample size increased and was highest when covariates were highly correlation (correlation = 0.9)
- Negative bias across all conditions, with smallest bias when covariates were highly correlated and highest when covariates were independent of each other (not expecting this result)
- BIC appears to be substantially influenced by correlation in the covariates
- MCP had higher proportion of correctly dropped covariates than SCAD in conditions in which only the penalty differed
- Next steps:
  - Why was bias smallest when covariates were highly correlated?
  - Compare penalties with empirical datasets to see if simulation condition results are corroborated