Don't be Fancy. Impute Your Dependent Variables!

Kyle M. Lang, Todd D. Little

Institute for Measurement, Methodology, Analysis & Policy
Texas Tech University
Lubbock, TX

May 24, 2016

Presented at the 6th Annual Modern Modeling Methods M^3 Conference
Storrs, CT

TEXAS TECH
UNIVERSITY

Outline

- Motivation and background
- Present simulation study
- Reiterate recommendations

Motivation

Pretty much everyone agrees that missing data should be treated with a principled analytic tool (i.e., FIML or MI).

Regression modeling offers an interesting special case.

- The basic regression problem is a relatively simple task.
- We only need to work with a single conditional density.
  - The predictors are usually assumed fixed.
- This simplicity means that many of the familiar problems with ad-hoc missing data treatments don't apply in certain regression modeling circumstances.

Special Case I

One familiar exception to the rule of always using a principled missing data treatment occurs when:

1. Missing data occur on the dependent variable of a linear regression model.
2. The missingness is strictly a function of the predictors in the regression equation.

Special Case I

One familiar exception to the rule of always using a principled missing data treatment occurs when:

1. Missing data occur on the dependent variable of a linear regression model.
2. The missingness is strictly a function of the predictors in the regression equation.

In this circumstance, listwise deletion (LWD) will produce unbiased estimates of the regression slopes.

- The intercept will be biased to the extent that missing data falls systematically closer to one tail of the DV's distribution.
- Power and generalizability still suffer from removing all cases that are subject to MAR missingness.

Complicating Special Case I

What if missing data occur on both the DV and IVs?

- Again, when missingness is strictly a function of IVs in the model, listwise deletion will produce unbiased estimates of regression slopes.
- If missingness on the IVs is a function of the DV, listwise deletion will bias slope estimates.
- Likewise when missingness is a function of unmeasured variables.

Complicating Special Case I

What if missing data occur on both the DV and IVs?

- Again, when missingness is strictly a function of IVs in the model, listwise deletion will produce unbiased estimates of regression slopes.
- If missingness on the IVs is a function of the DV, listwise deletion will bias slope estimates.
- Likewise when missingness is a function of unmeasured variables.

When missingness occurs on both the DV and IVs, the general recommendation is to use MI to impute all missing data.

- Little (1992) showed that including the incomplete DV in the imputation model can improve imputations of the IVs.

Special Case II

There is still debate about how to address the cases with imputed DV values.

- Von Hippel (2007) introduced the *Multiple Imputation then Deletion* (MID) approach.
  - Von Hippel (2007) claimed that cases with imputed DV values cannot provide any information to the regression equation.
  - He suggested that such cases should be retained for imputation but should be excluded from the final inferential modeling.
  - Von Hippel (2007) provided analytic and simulation-based arguments for the superiority of MID to traditional MI (wherein the imputed DVs are retained for inferential analyses).

Rationale for MID

The MID approach rests on the following premises:

1. Observations with missing DVs cannot offer any information to the estimation of regression slopes.
2. Including these observations can only increase the between-imputation variability of the pooled estimates.

Rationale for MID

The MID approach rests on the following premises:

1. Observations with missing DVs cannot offer any information to the estimation of regression slopes.
2. Including these observations can only increase the between-imputation variability of the pooled estimates.

BUT, there are a two big issues with this foundation:

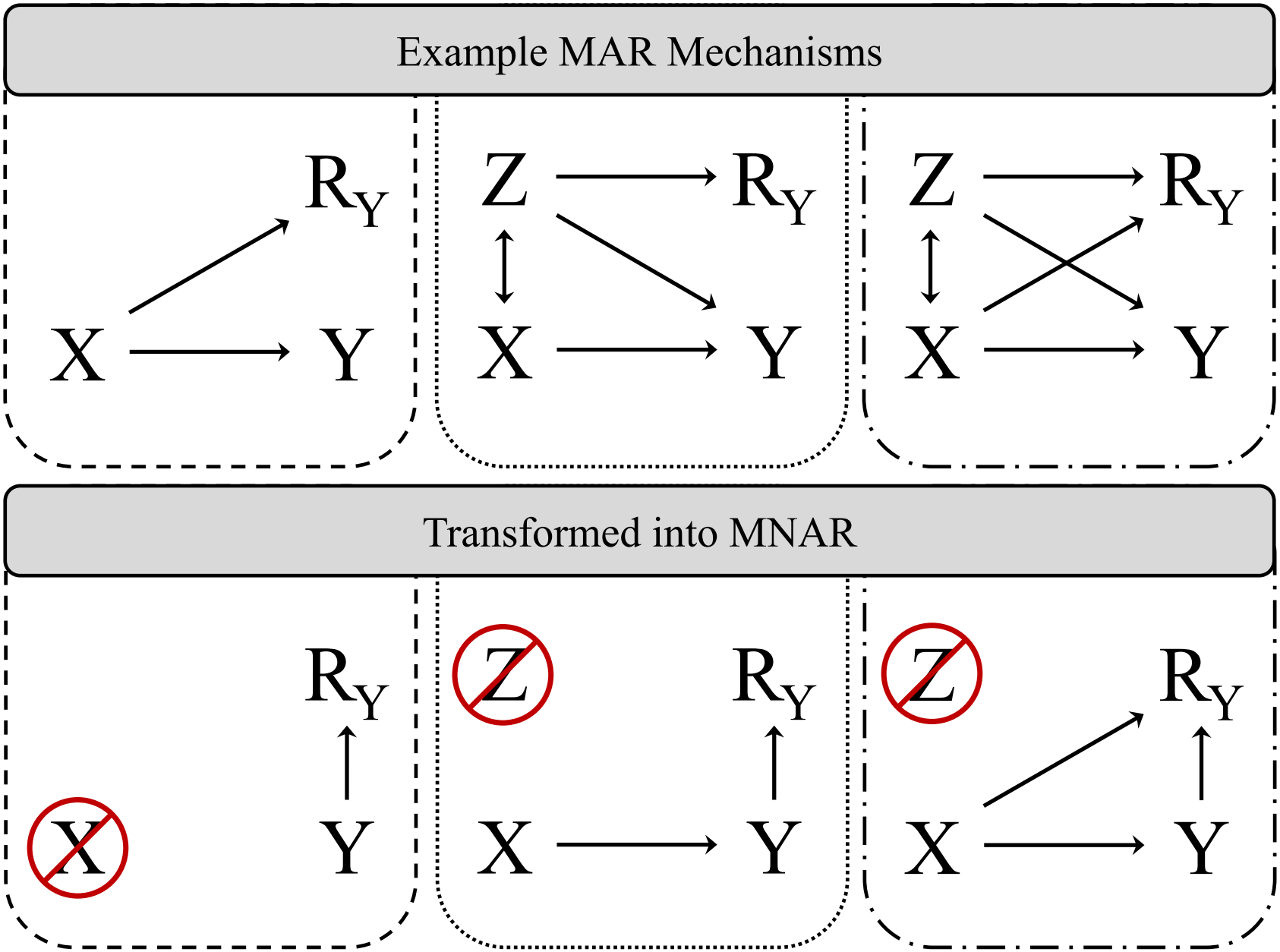
1. Premise 1 is only true when the MAR predictors are fully represented among the IVs of the inferential regression model.
2. Premise 2 is nullified by taking a large enough number of imputations.

Crux of the Matter

This whole problem boils down to whether or not the MAR assumption is satisfied in the inferential model.

- Special Cases I and II amount to situations wherein the inferential regression model suffices to satisfy the MAR assumption.
- In general, neither LWD nor MID will satisfy the MAR assumption.
- When any portion of the (multivariate) MAR predictor is not contained by the set of IVs in the inferential model, both LWD and MID will produce biased estimates of regression slopes.
  - Given the minor caveat I'll discuss momentarily

Graphical Representations



Methods

Simulation Parameters

Primary parameters

1. Proportion of the (bivariate) MAR predictor that was represented among the analysis model's IVs:
   - $pMAR = \{1.0, 0.75, 0.5, 0.25, 0.0\}$
2. Strength of correlations among the predictors in the data generating model:
   - $rXZ = \{0.0, 0.1, 0.3, 0.5\}$

Simulation Parameters

Primary parameters

1. Proportion of the (bivariate) MAR predictor that was represented among the analysis model's IVs:
   - $pMAR = \{1.0, 0.75, 0.5, 0.25, 0.0\}$
2. Strength of correlations among the predictors in the data generating model:
   - $rXZ = \{0.0, 0.1, 0.3, 0.5\}$

Secondary parameters

- Sample size: $N = \{100, 250, 500\}$
- Proportion of missing data: $PM = \{0.1, 0.2, 0.4\}$
- R^2 for the data generating model: $R^2 = \{0.15, 0.3, 0.6\}$

Simulation Parameters

Primary parameters

1. Proportion of the (bivariate) MAR predictor that was represented among the analysis model's IVs:
   - $pMAR = \{1.0, 0.75, 0.5, 0.25, 0.0\}$
2. Strength of correlations among the predictors in the data generating model:
   - $rXZ = \{0.0, 0.1, 0.3, 0.5\}$

Secondary parameters

- Sample size: $N = \{100, 250, 500\}$
- Proportion of missing data: $PM = \{0.1, 0.2, 0.4\}$
- R^2 for the data generating model: $R^2 = \{0.15, 0.3, 0.6\}$

Crossed conditions in the final design

- $5(pMAR) \times 4(rXZ) \times 3(N) \times 3(PM) \times 3(R^2) = 540$
- $R = 500$ replications within each condition.

Data Generation

Data were generated according to the following model:

$$Y = 1.0 + 0.33X + 0.33Z_1 + 0.33Z_2 + \varepsilon,$$
$$\varepsilon \sim \mathrm{N}\left(0, \sigma^2\right).$$

Where σ^2 was manipulated to achieve the desired R^2 level.

Data Generation

Data were generated according to the following model:

$$Y = 1.0 + 0.33X + 0.33Z_1 + 0.33Z_2 + \varepsilon,$$
$$\varepsilon \sim \mathrm{N}\left(0, \sigma^2\right).$$

Where σ^2 was manipulated to achieve the desired R^2 level.

The analysis model was: $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X + \hat{\beta}_2 Z_1$.

Data Generation

Data were generated according to the following model:

$$Y = 1.0 + 0.33X + 0.33Z_1 + 0.33Z_2 + \varepsilon,$$
$$\varepsilon \sim \mathrm{N}\left(0, \sigma^2\right).$$

Where σ^2 was manipulated to achieve the desired R^2 level.

The analysis model was: $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X + \hat{\beta}_2 Z_1$.

Missing data were imposed on Y and X using the weighted sum of Z_1 and Z_2 as the MAR predictor.

- The weighting was manipulated to achieve the proportions of MAR in $\{pMAR\}$.
- Y values in the positive tail of the MAR predictor's distribution and X values in the negative tail of the MAR predictor's distribution were set to missing data.

Outcome Measures

The focal parameter was the slope coefficient associated with X in the analysis model (i.e., β_1).

For this report, we focus on two outcome measures:

1. Percentage Relative Bias:

$$PRB = 100 \times \frac{\bar{\hat{\beta}}_1 - \beta_1}{\beta_1}$$

2. Empirical Power:

$$Power = R^{-1} \sum_{r=1}^{R} \mathrm{I}\left(p_{\beta 1, r} < 0.05\right)$$

- True values (i.e., β_1) were the average complete data estimates.

Computational Details

The simulation code was written in the R statistical programming language (R Core Team, 2014).

Missing data were imputed using the **mice** package (van Buuren & Groothuis-Oudshoorn, 2011).

- $m = 100$ imputations were created.

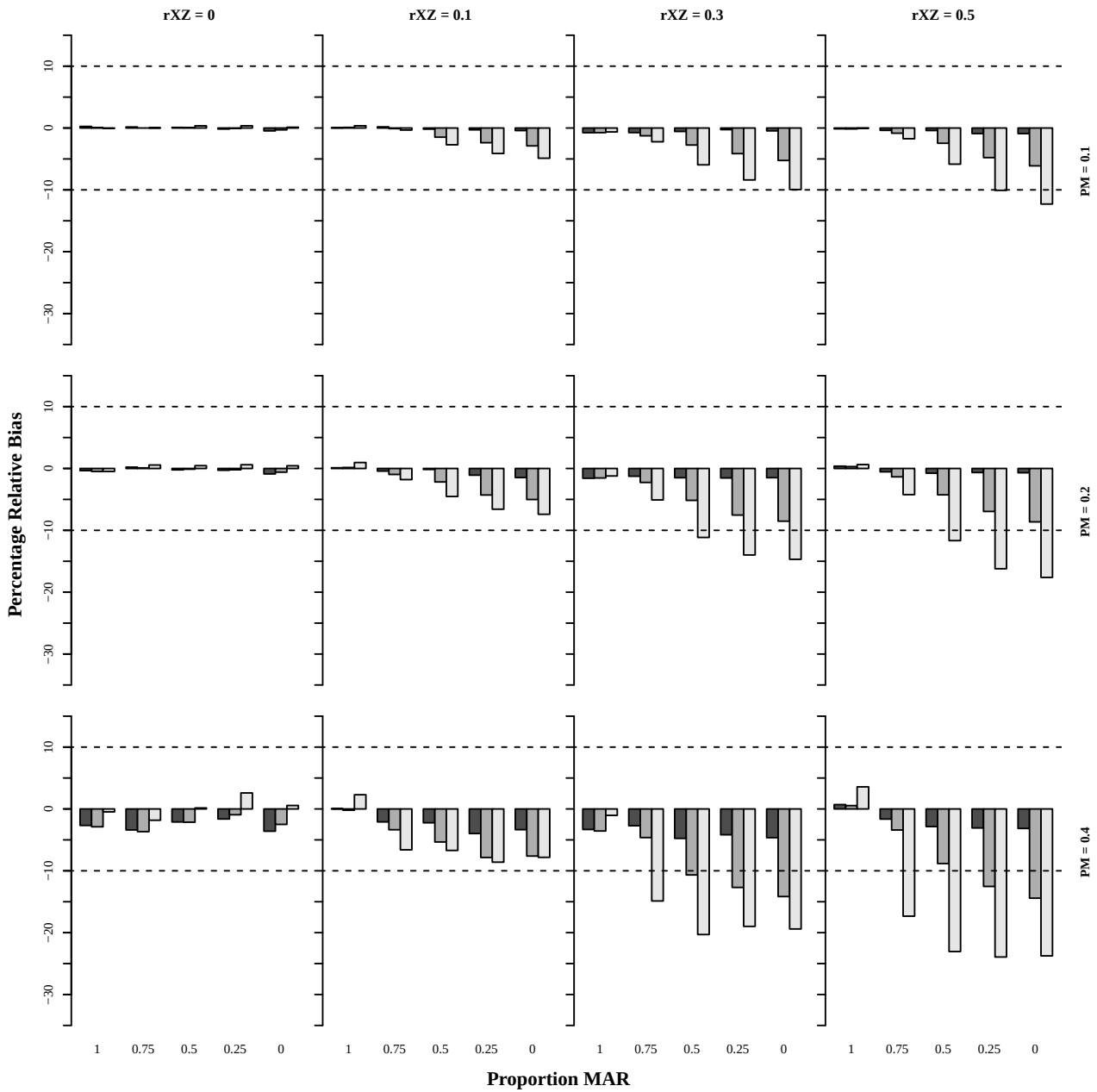
Results were pooled using the **mitools** package (Lumley, 2014).

Hypotheses

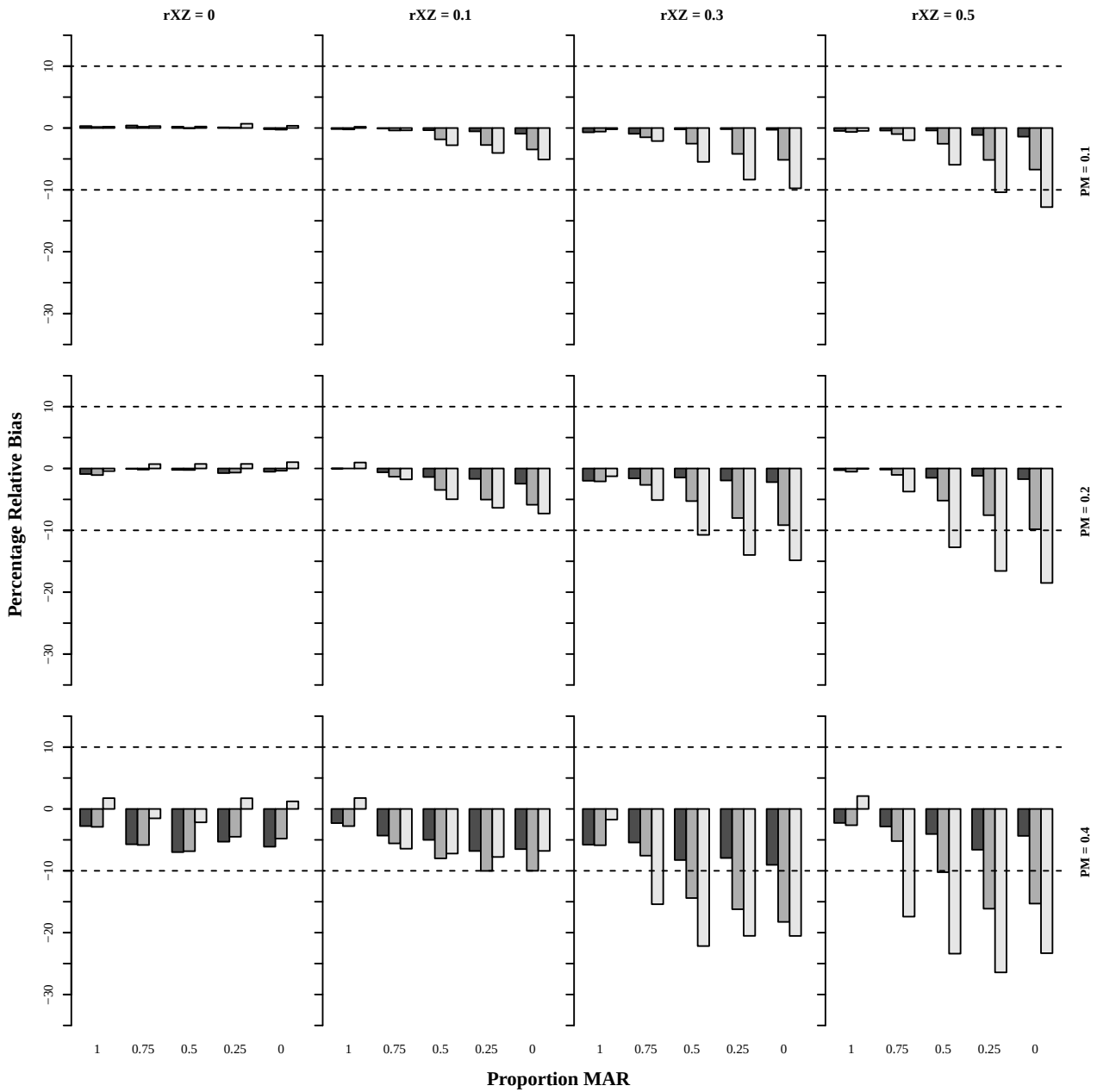
1. Traditional MI will produce unbiased estimates of β_1 in all conditions.
2. When $rXZ = 0.0$ or $pMAR = 1.0$, MID and LWD will produce unbiased estimates of β_1.
3. When $pMAR \neq 1.0$ and $rXZ \neq 0.0$, MID and LWD will produce biased estimates of β_1 and bias will increase as $pMAR$ decreases and rXZ increases.
4. Traditional MI will maintain power levels that are, at least, as high as MID and LWD in all conditions.
5. LWD and MID will manifest disproportionately greater power loss than traditional MI.

Results

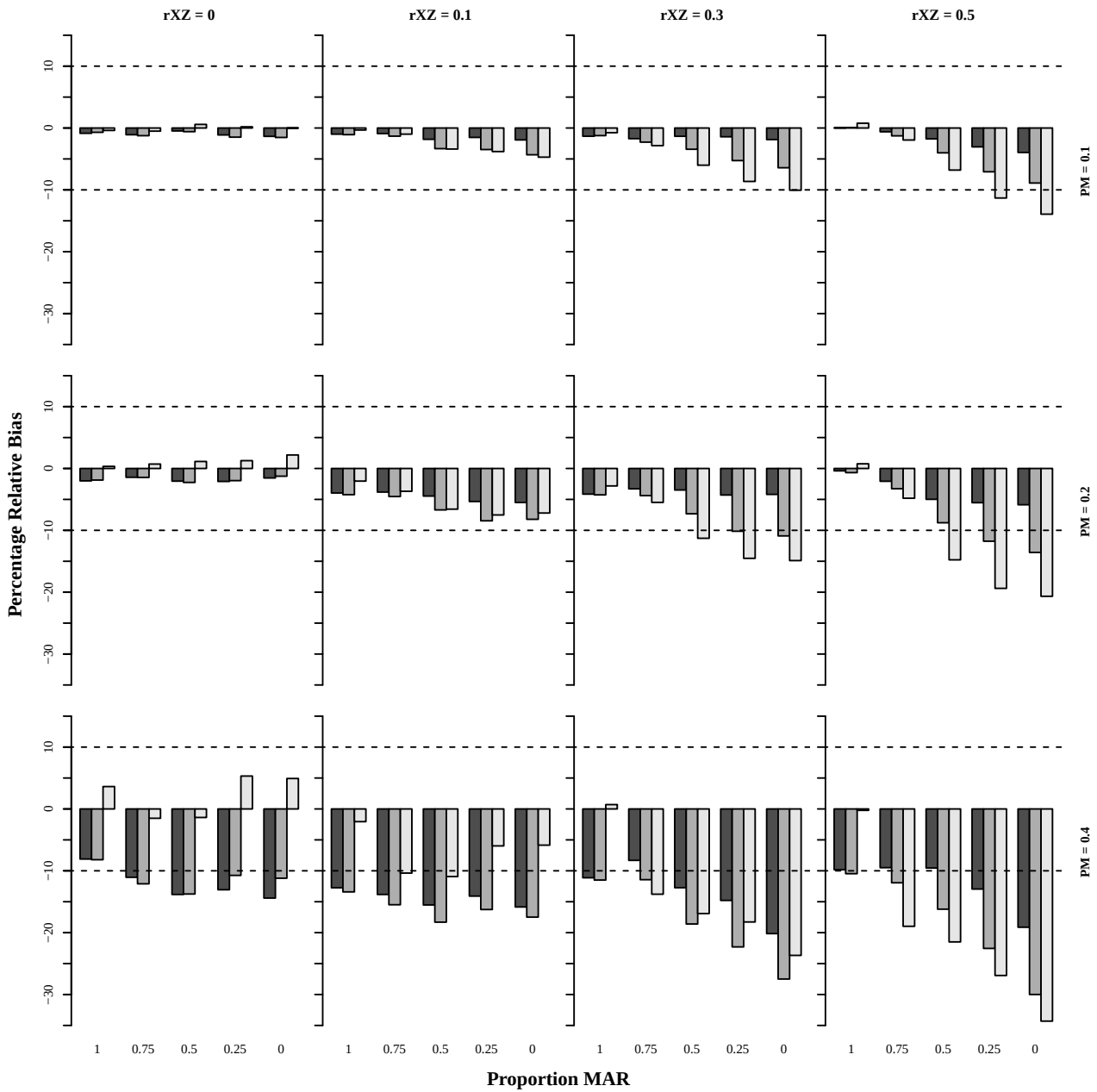
PRB: $N = 500$; $R^2 = 0.3$



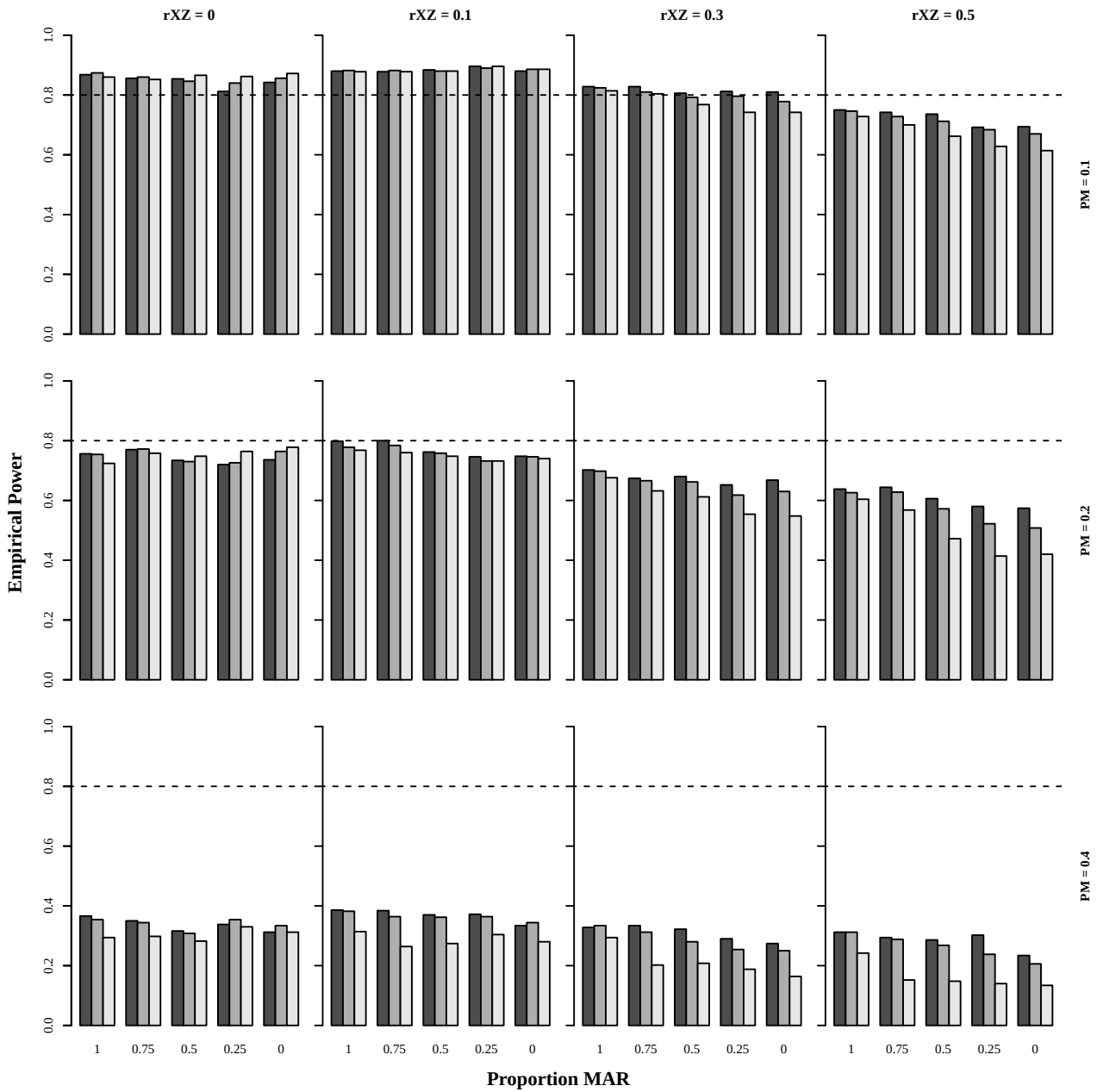
PRB: $N = 250$; $R^2 = 0.3$



PRB: $N = 100$; $R^2 = 0.3$



Power: $N = 100$; $R^2 = 0.3$



Discussion

Bias-Related Findings

Traditional MI did not lead to bias in most conditions.

- When $N = 100$ and $PM = 0.4$ all methods induced relatively large biases.

Bias-Related Findings

Traditional MI did not lead to bias in most conditions.

- When $N = 100$ and $PM = 0.4$ all methods induced relatively large biases.

Unless the set of MAR predictors is a subset of the IVs in the analysis model, MID and LWD will produce biased estimates of regression slopes.

- Traditional MI requires only that the MAR predictors be available for use during the imputation process.

Bias-Related Findings

Traditional MI did not lead to bias in most conditions.

- When $N = 100$ and $PM = 0.4$ all methods induced relatively large biases.

Unless the set of MAR predictors is a subset of the IVs in the analysis model, MID and LWD will produce biased estimates of regression slopes.

- Traditional MI requires only that the MAR predictors be available for use during the imputation process.

Science prefers parsimonious models, so it seems likely that important MAR predictors are often not represented in the set of analyzed IVs.

Power-Related Findings

Traditional MI did not suffer greater power loss than MID or LWD.

- Taking sufficiently many imputations mitigates any inflation of variability due to between-imputation variance.
- Arguments for MI's inflation of variability are all based on use of a very small number of imputations
- The commonly cited justification for few (i.e., $m = 5$) imputations was made in 1987 (i.e., Rubin, 1987).

Power-Related Findings

Traditional MI did not suffer greater power loss than MID or LWD.

- Taking sufficiently many imputations mitigates any inflation of variability due to between-imputation variance.
- Arguments for MI's inflation of variability are all based on use of a very small number of imputations
- The commonly cited justification for few (i.e., $m = 5$) imputations was made in 1987 (i.e., Rubin, 1987).

Both MID and LWD suffered substantial power loss with high proportions of missing data

- No matter the mathematical justification, both MID and LWD entail throwing away large portions of your dataset.

Conclusion

In special circumstances, LWD and MID will produce unbiased estimates of regression slopes, but...

- These conditions are not likely to occur outside of strictly controlled experimental settings.
- The negative consequences of assuming these special conditions hold, when they do not, can be severe.
- Estimated intercepts, means, variances, and correlations will still be biased.

Conclusion

The only methodological argument against traditional MI in favor of MID assumes the use of a very small number of imputations (i.e., $m < 10$).

- Taking m to be large enough ensures that traditional MI will do no worse than MID.
- Traditional MI will perform well if the MAR predictors are available, without required them to be included in the analysis model.

Limitations

The models we employed were very simple.

- Some may question the ecological validity of our results.
- We purposefully focused on internal validity.

Limitations

The models we employed were very simple.

- Some may question the ecological validity of our results.
- We purposefully focused on internal validity.

All relationships were linear.

- The findings presented here may not fully generalize to:
  - MAR mechanisms that manifest through nonlinear relations.
  - Nonlinear regression models (e.g., moderated regression, polynomial regression, generalized linear models).
- These nonlinear situations are important areas for future work.

Take Home Message

Don't be Fancy.

Impute your DVs!

(and don't delete them, afterward)

References

Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, *87*(420), 1227–1237.

Lumley, T. (2014). mitools: Tools for multiple imputation of missing data [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=mitools` (R package version 2.3)

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

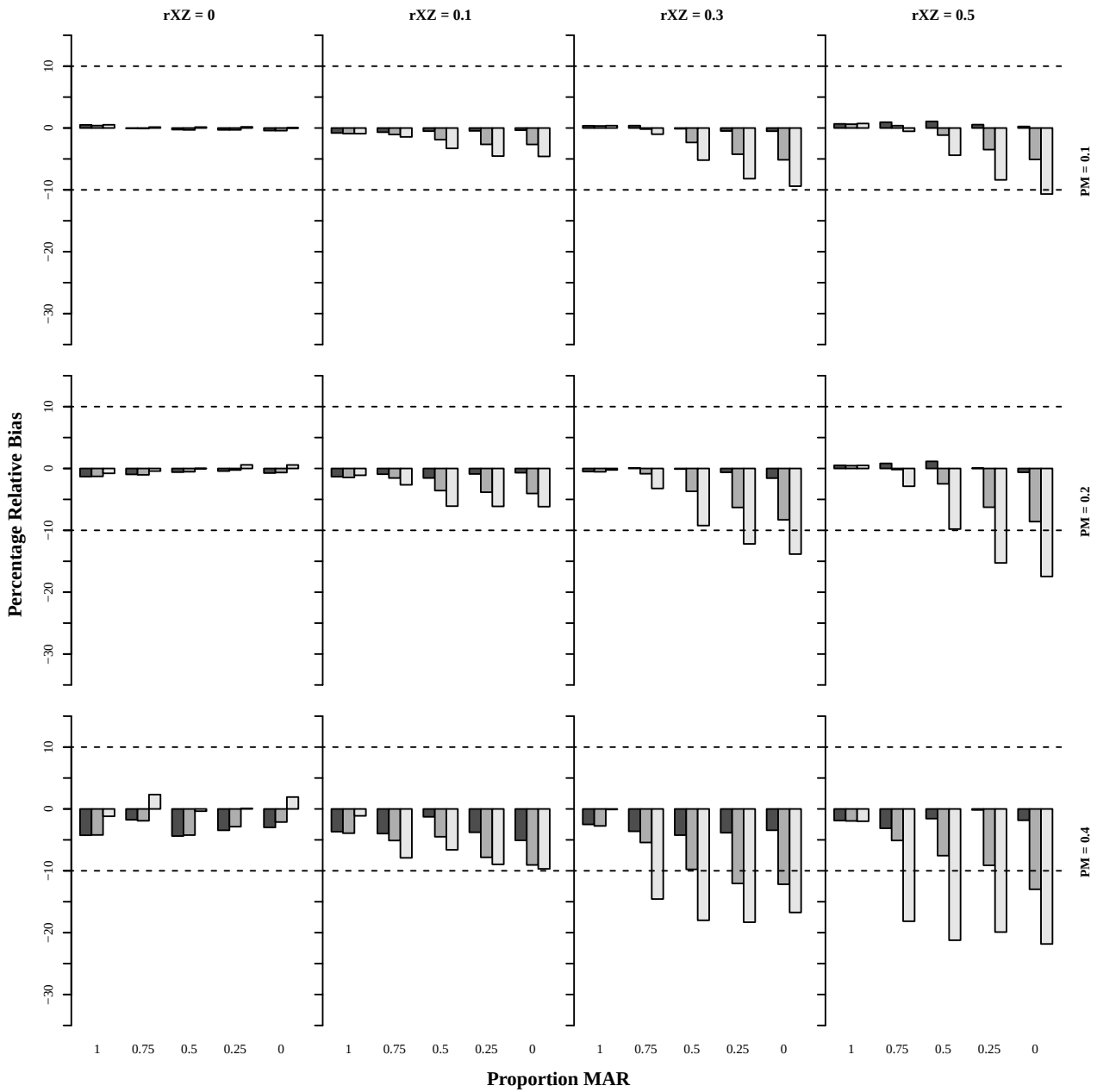
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (Vol. 519). New York, NY: John Wiley & Sons.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

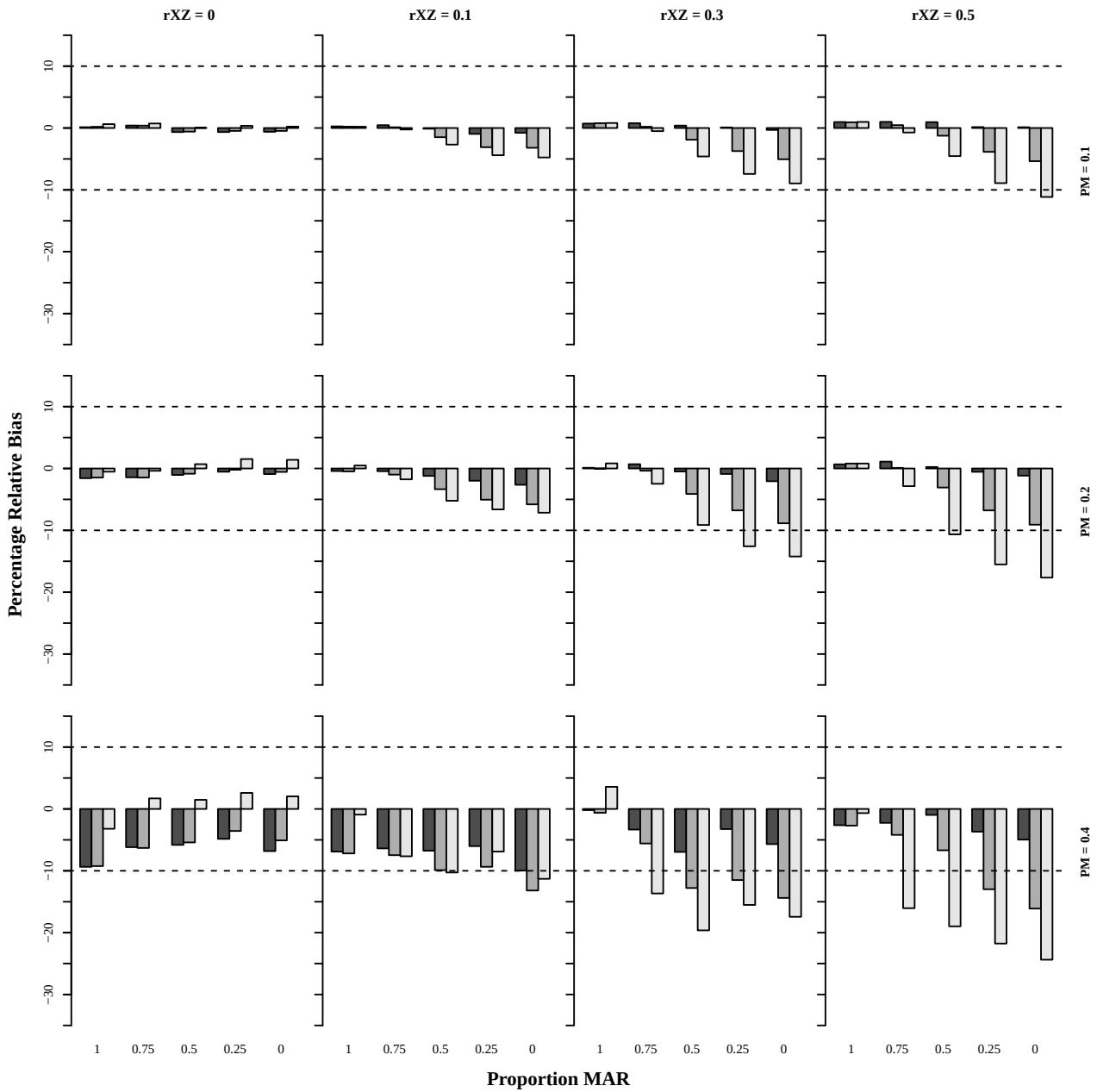
Von Hippel, P. T. (2007). Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, *37*(1), 83–117.

Extras

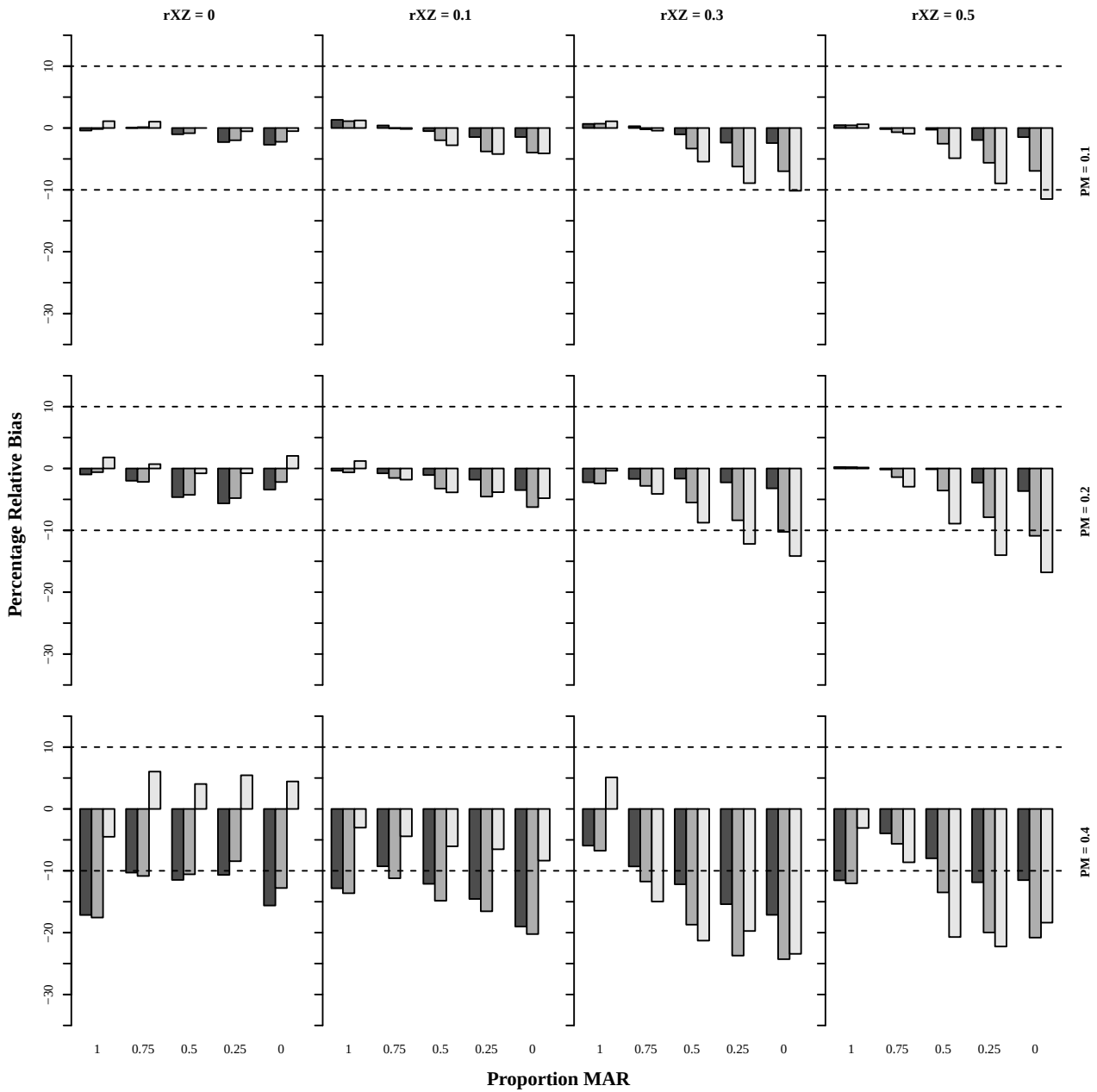
PRB: $N = 500$; $R^2 = 0.15$



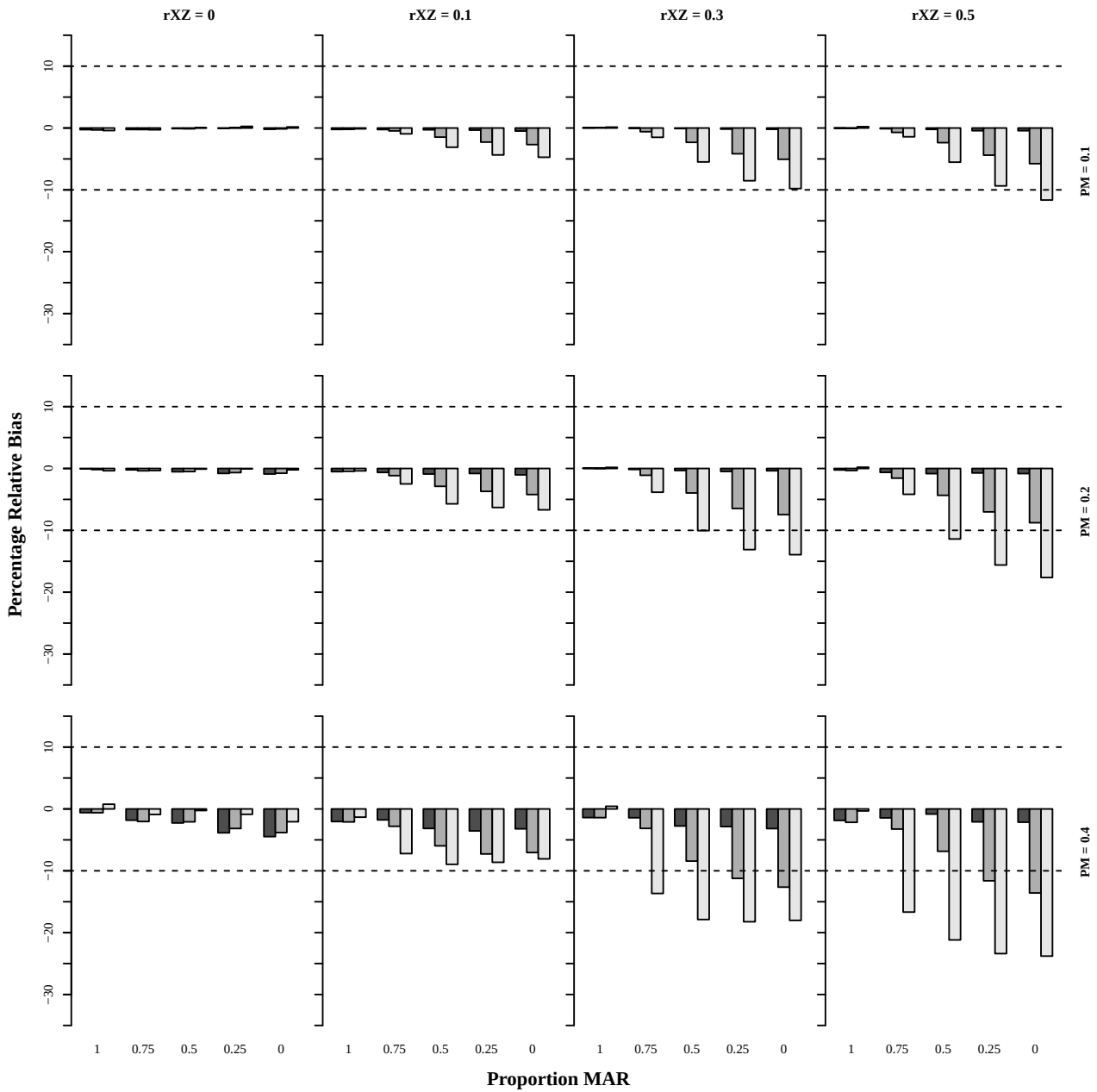
PRB: $N = 250$; $R^2 = 0.15$



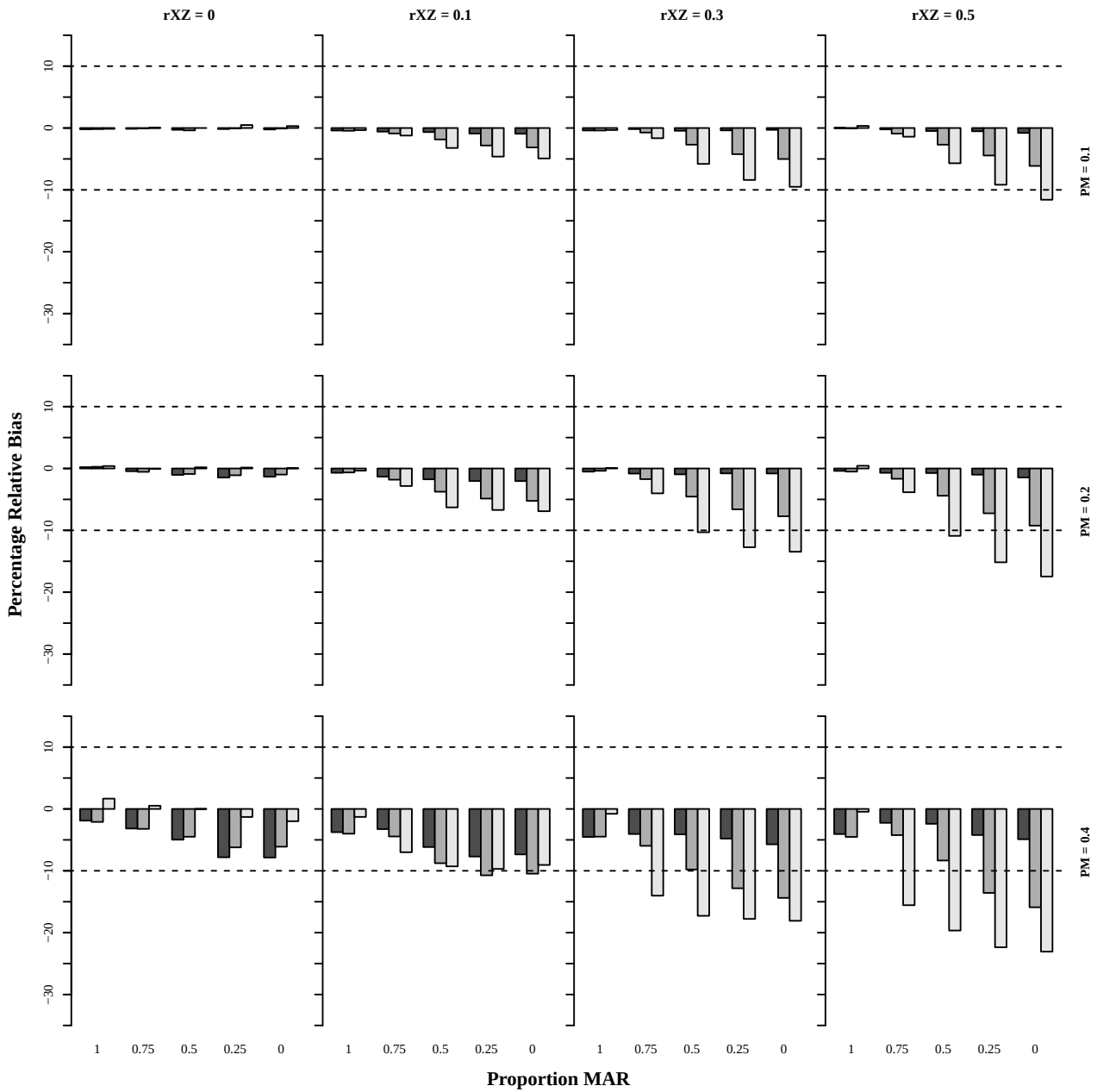
PRB: $N = 100$; $R^2 = 0.15$



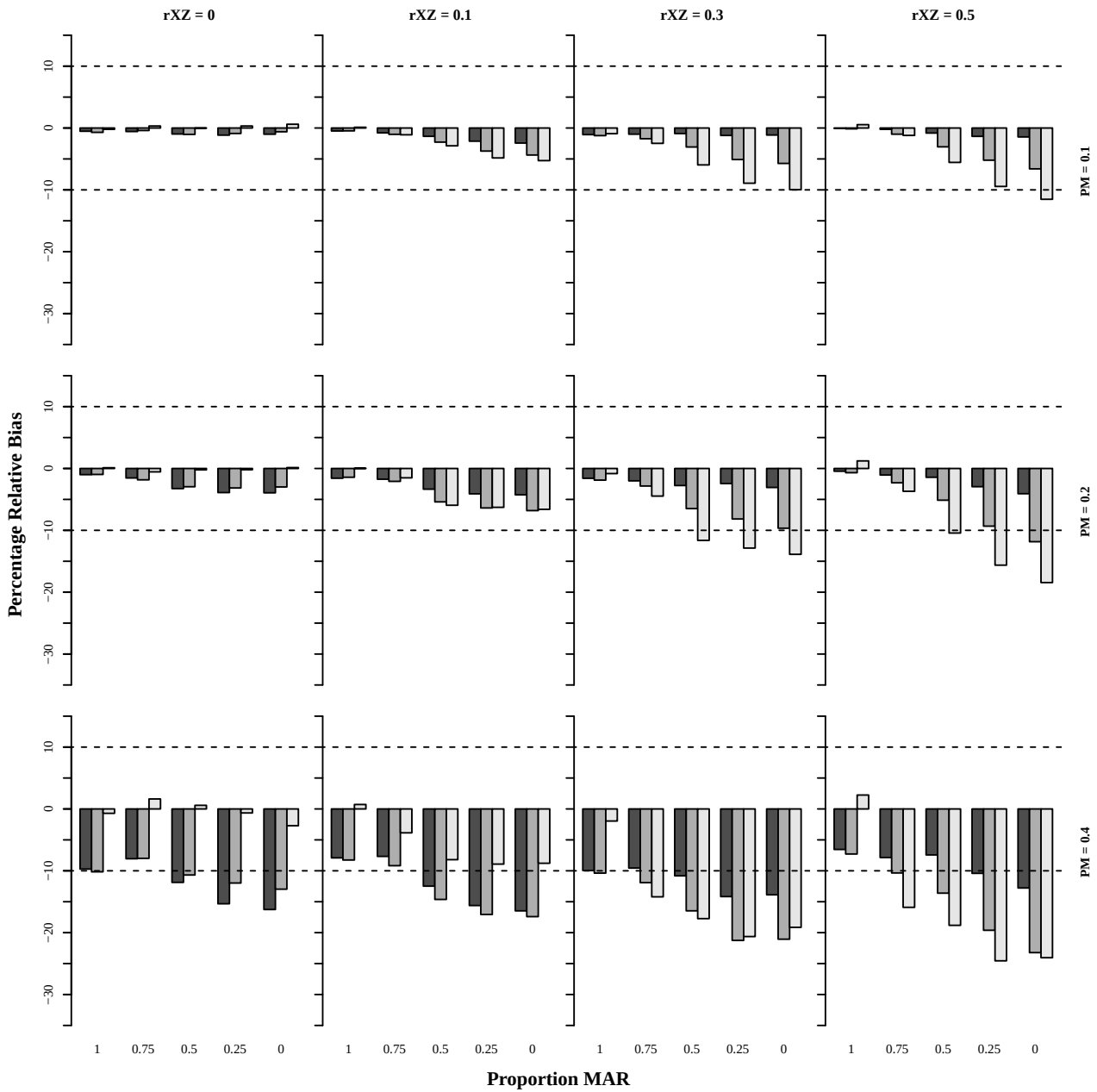
PRB: $N = 500$; $R^2 = 0.60$



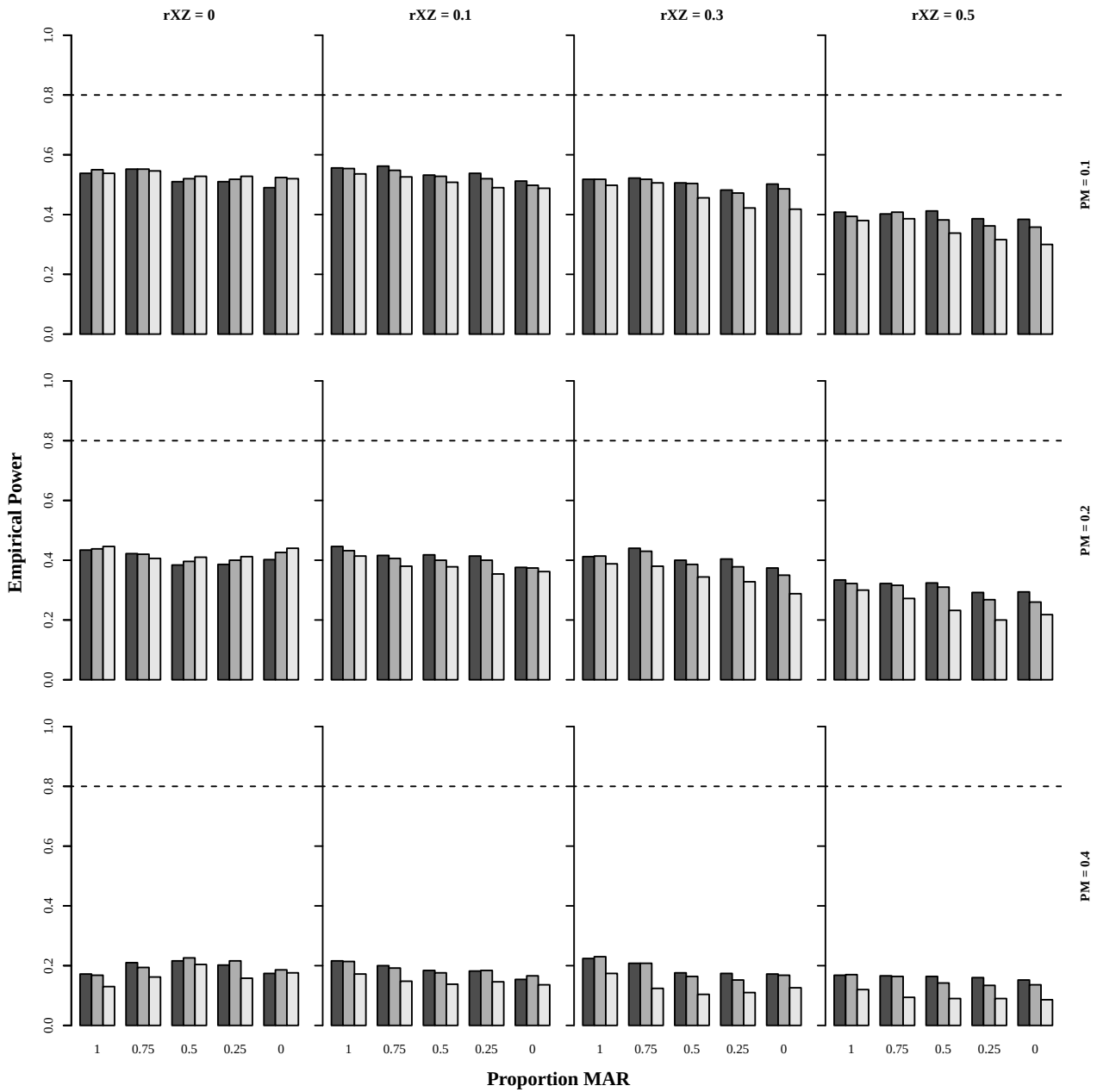
PRB: $N = 250$; $R^2 = 0.60$



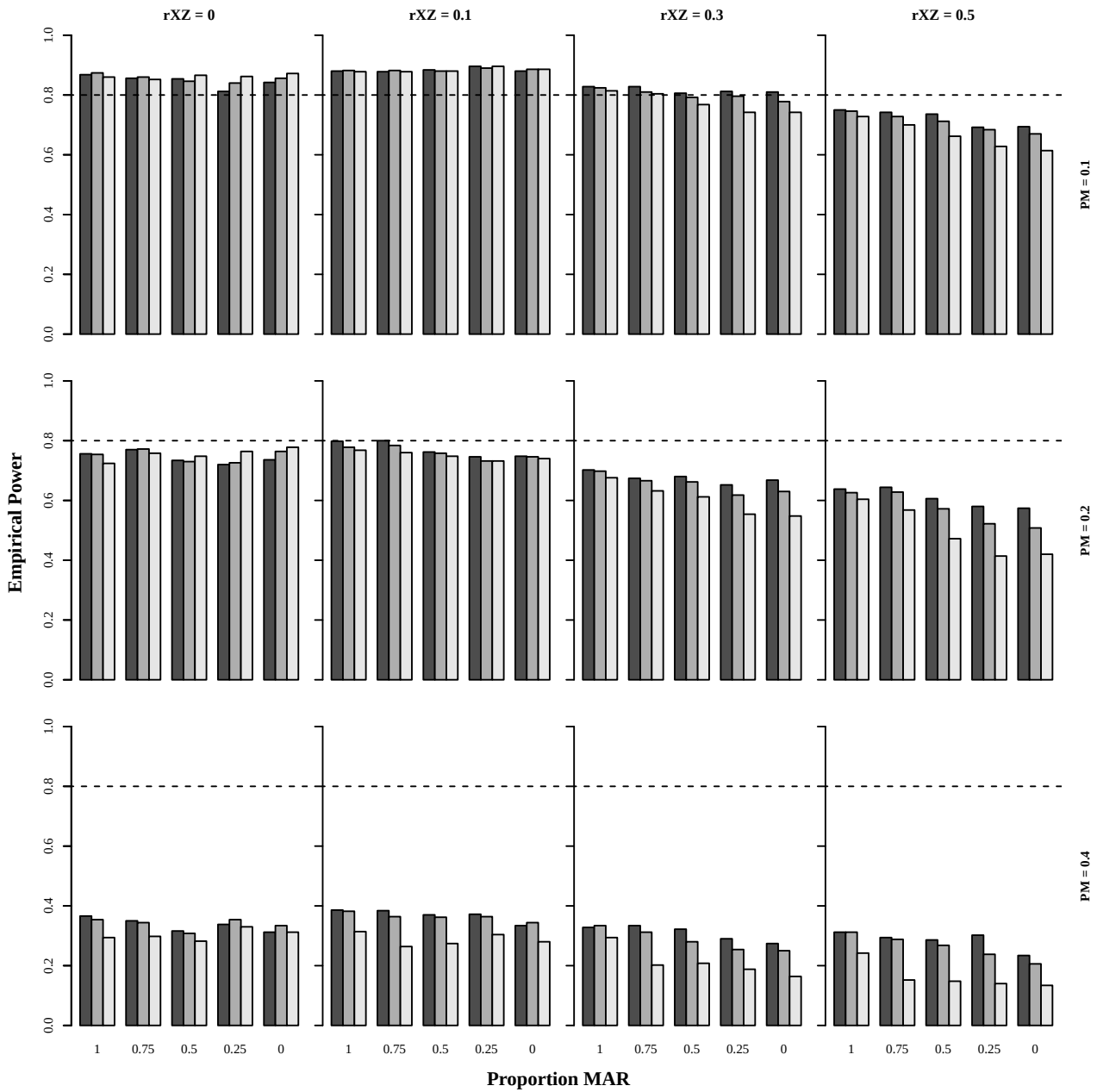
PRB: $N = 100$; $R^2 = 0.60$



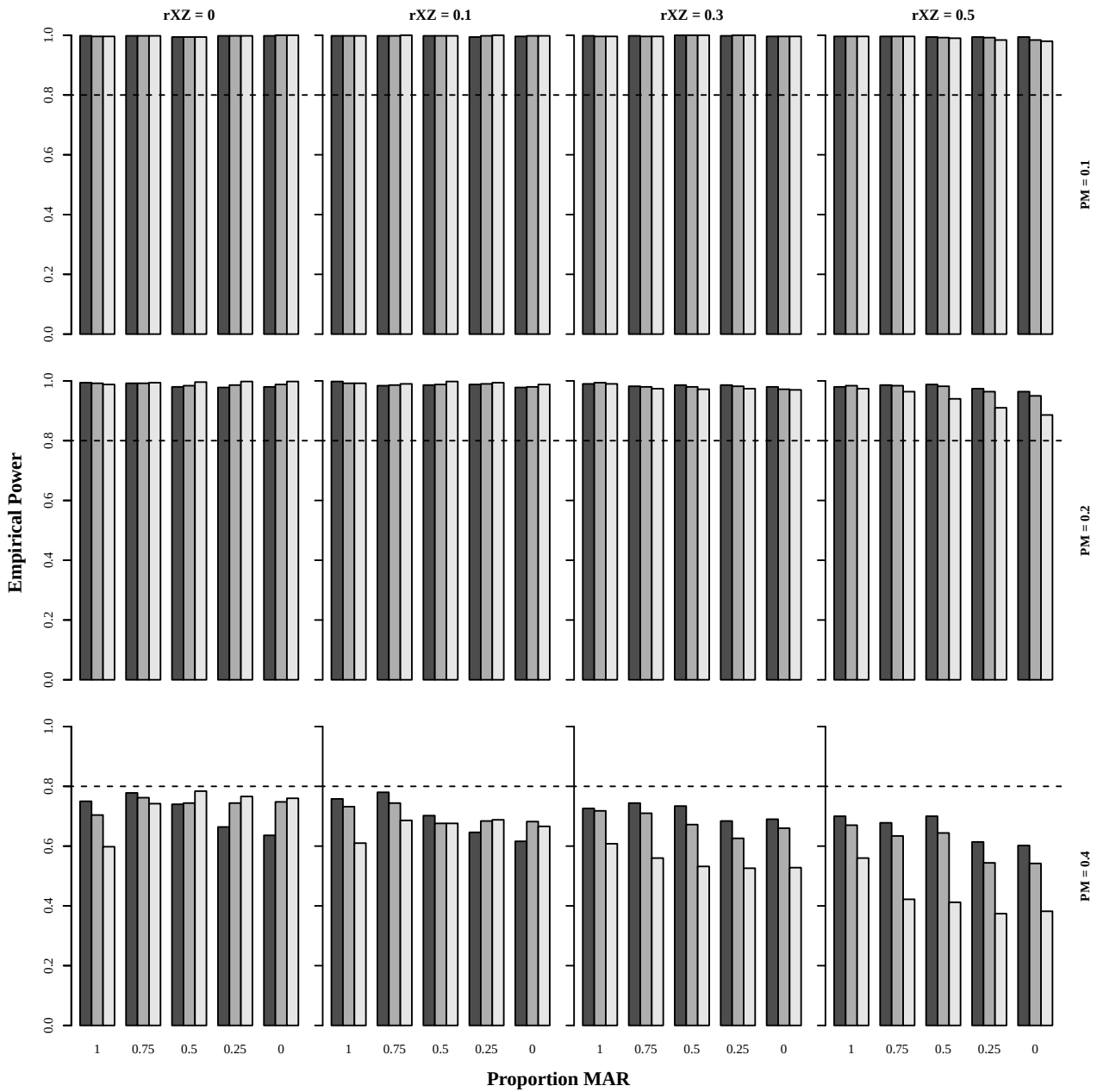
Power: $N = 100$; $R^2 = 0.15$



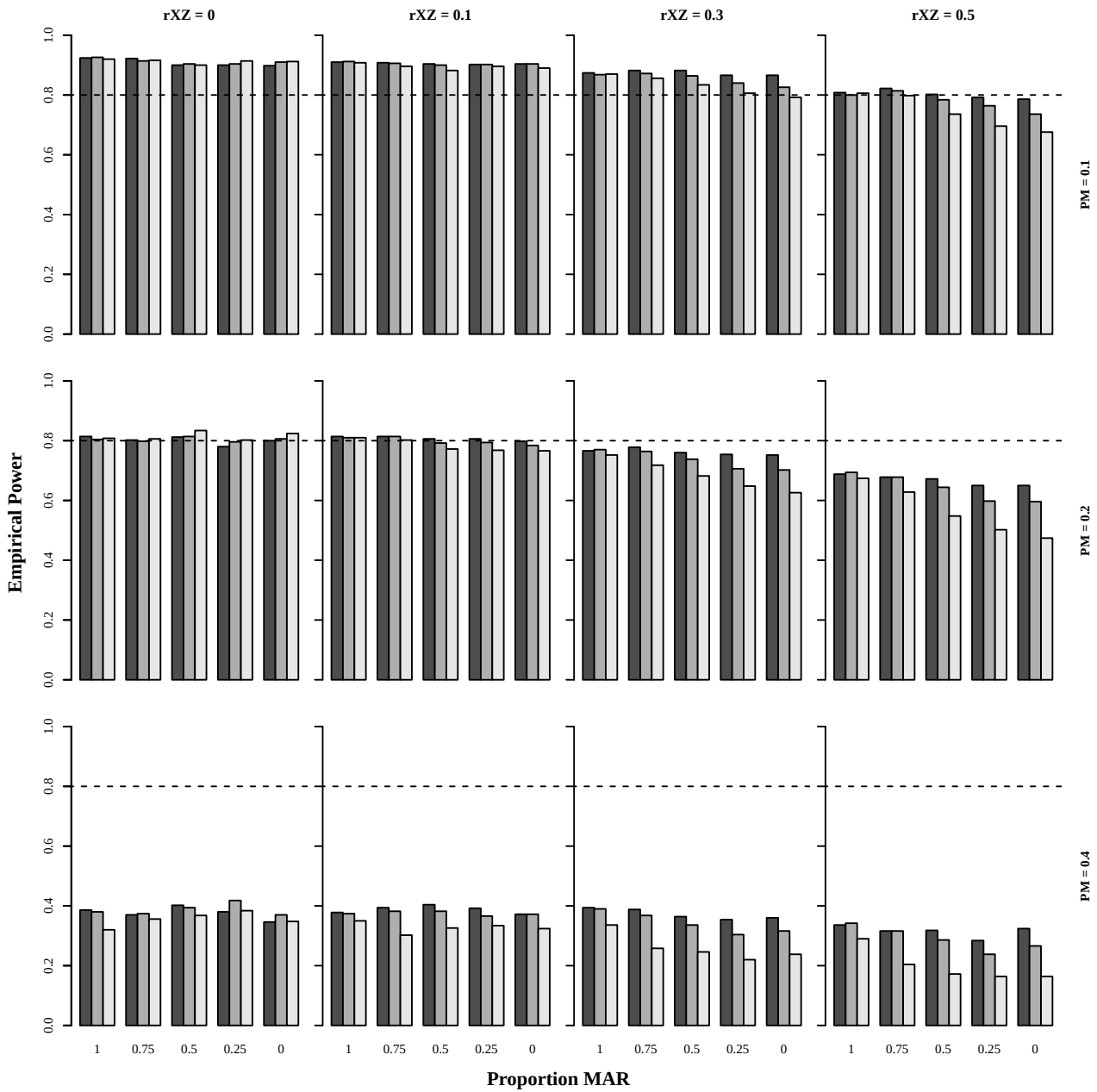
Power: $N = 100$; $R^2 = 0.3$



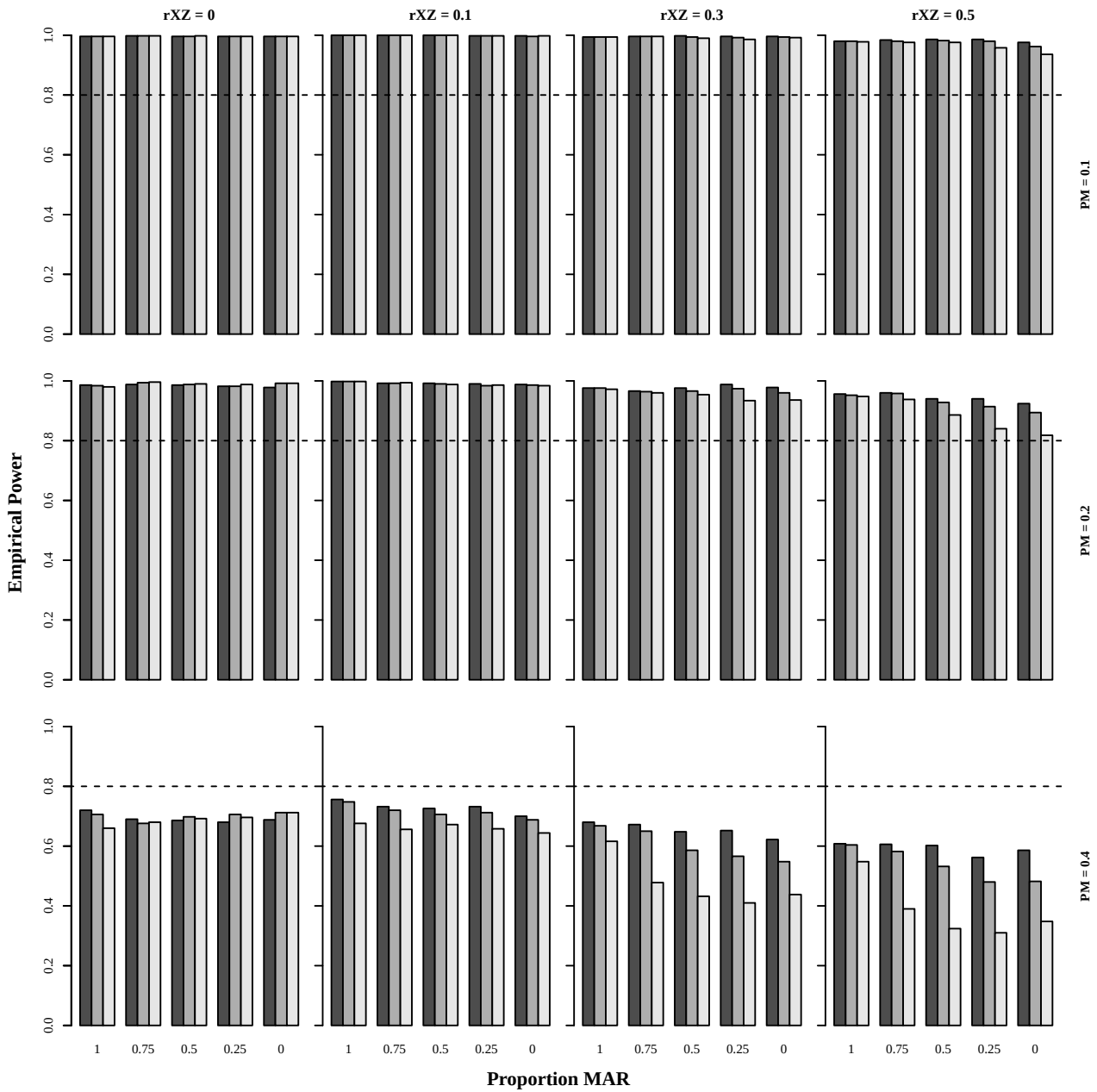
Power: $N = 100$; $R^2 = 0.6$



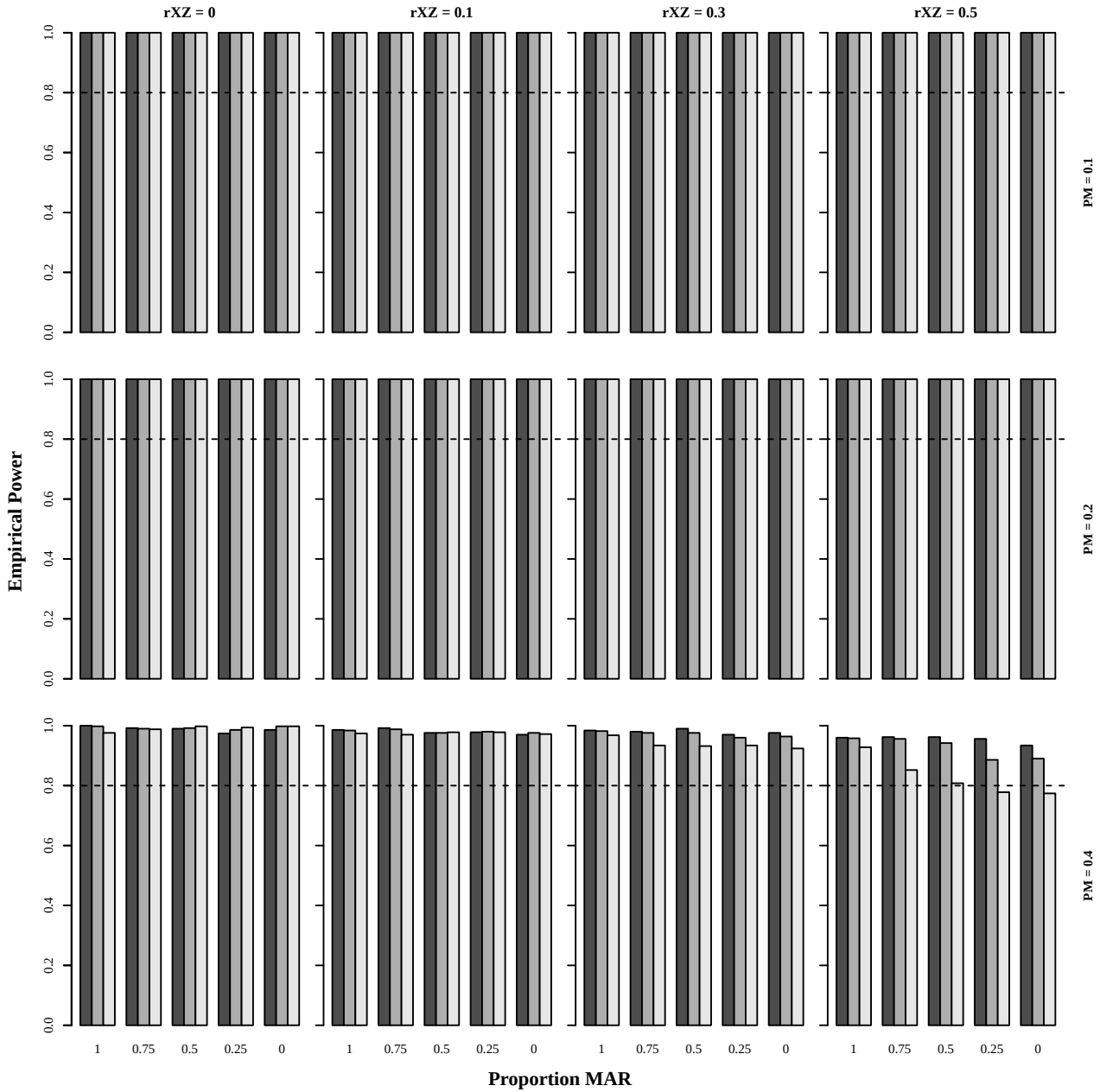
Power: $N = 250$; $R^2 = 0.15$



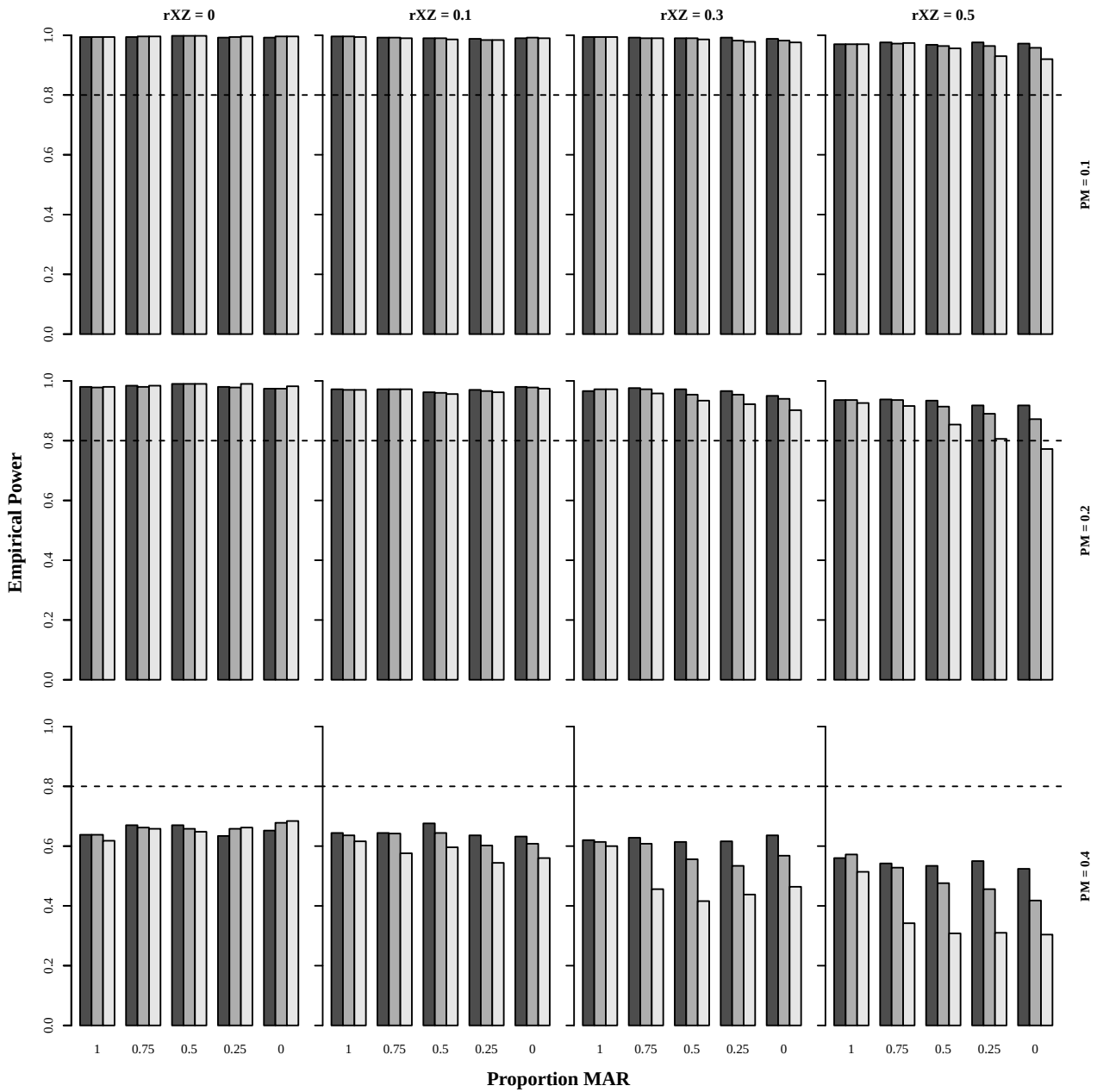
Power: $N = 250$; $R^2 = 0.30$



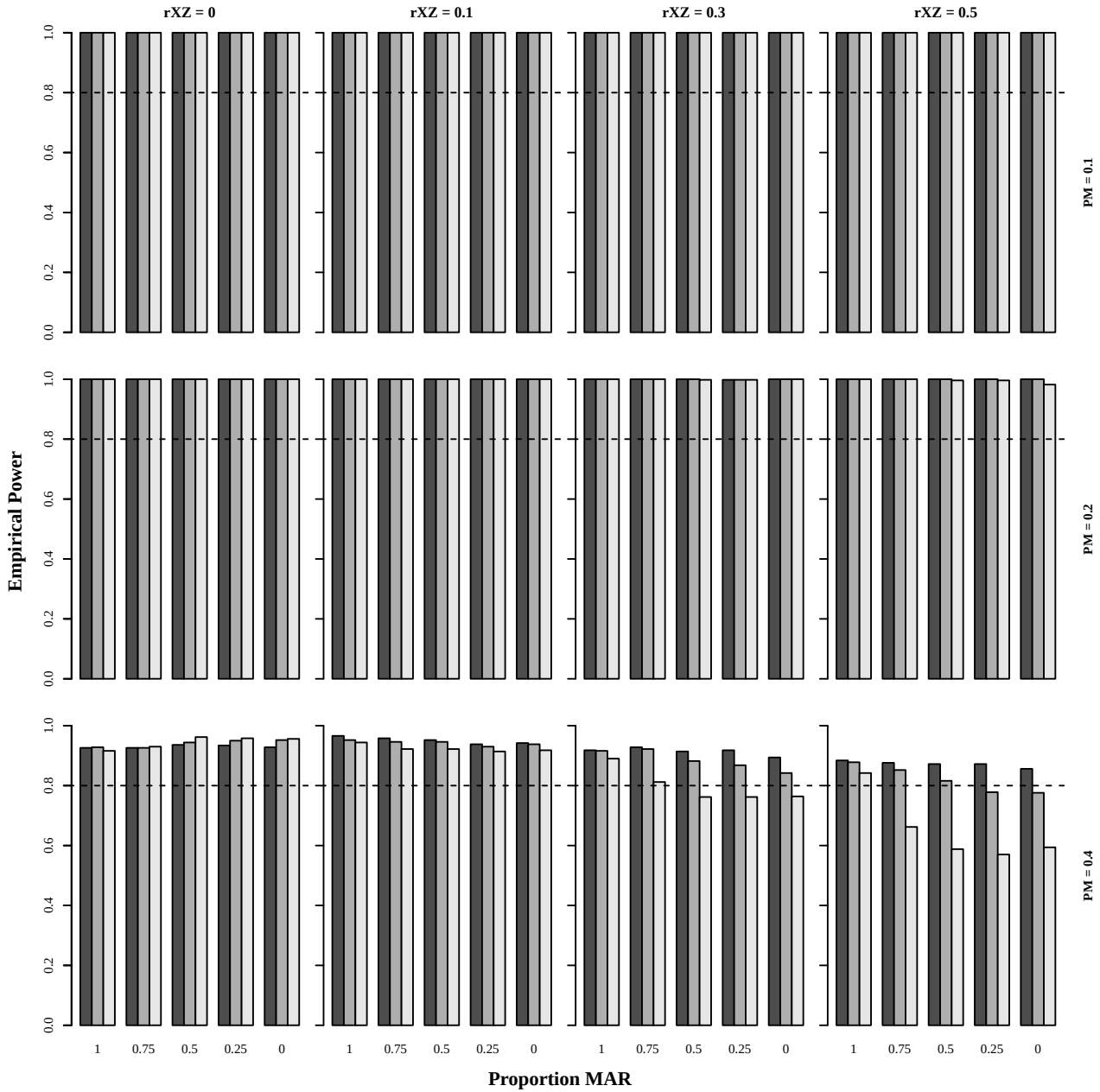
Power: $N = 250$; $R^2 = 0.60$



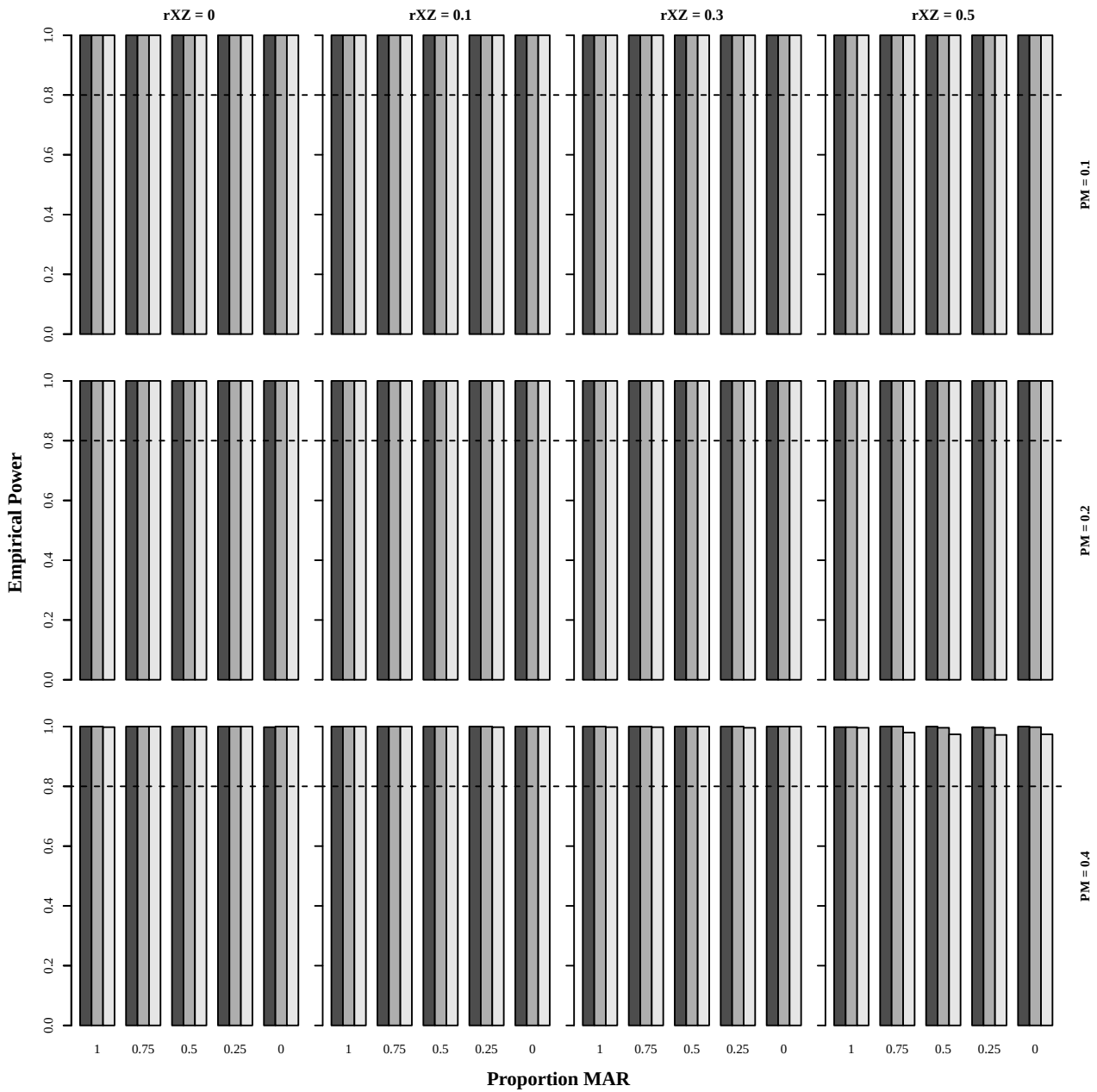
Power: $N = 500$; $R^2 = 0.15$



Power: $N = 500$; $R^2 = 0.30$



Power: $N = 500$; $R^2 = 0.60$



Influence of Von Hippel (2007)

The MID technique has become relatively popular.

- A *Web of Science* search for citations of Von Hippel (2007) returns 297 results.
- Filtering those hits to only psychology related subjects results in 79 citations.
- Of these 79 papers, 60 (75.95%) employed the MID approach in empirical research.