Should You Impute Incomplete Dependent Variables in Multiple Linear Regression Models?

Kyle M. Lang

Department of Methodology & Statistics
Tilburg University

October 17, 2017

Outline

- Motivation and background
- Present two simulation studies
- 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 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

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

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

Special Case

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

1. Missing data occur only on the dependent variable.
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.

General Case

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

- 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.

General Case

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

- 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.

Treating the General Case

Treatment of cases with imputed DV values continues to spur debate.

- Von Hippel (2007) introduced the *Multiple Imputation then Deletion* (MID) approach.
  - Claimed that cases with imputed DV values cannot provide any information to the regression equation.
  - Suggested that such cases should be retained for imputation but should be excluded from the final inferential modeling.
  - 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.

Influence of Von Hippel (2007)

The MID technique has become relatively popular.

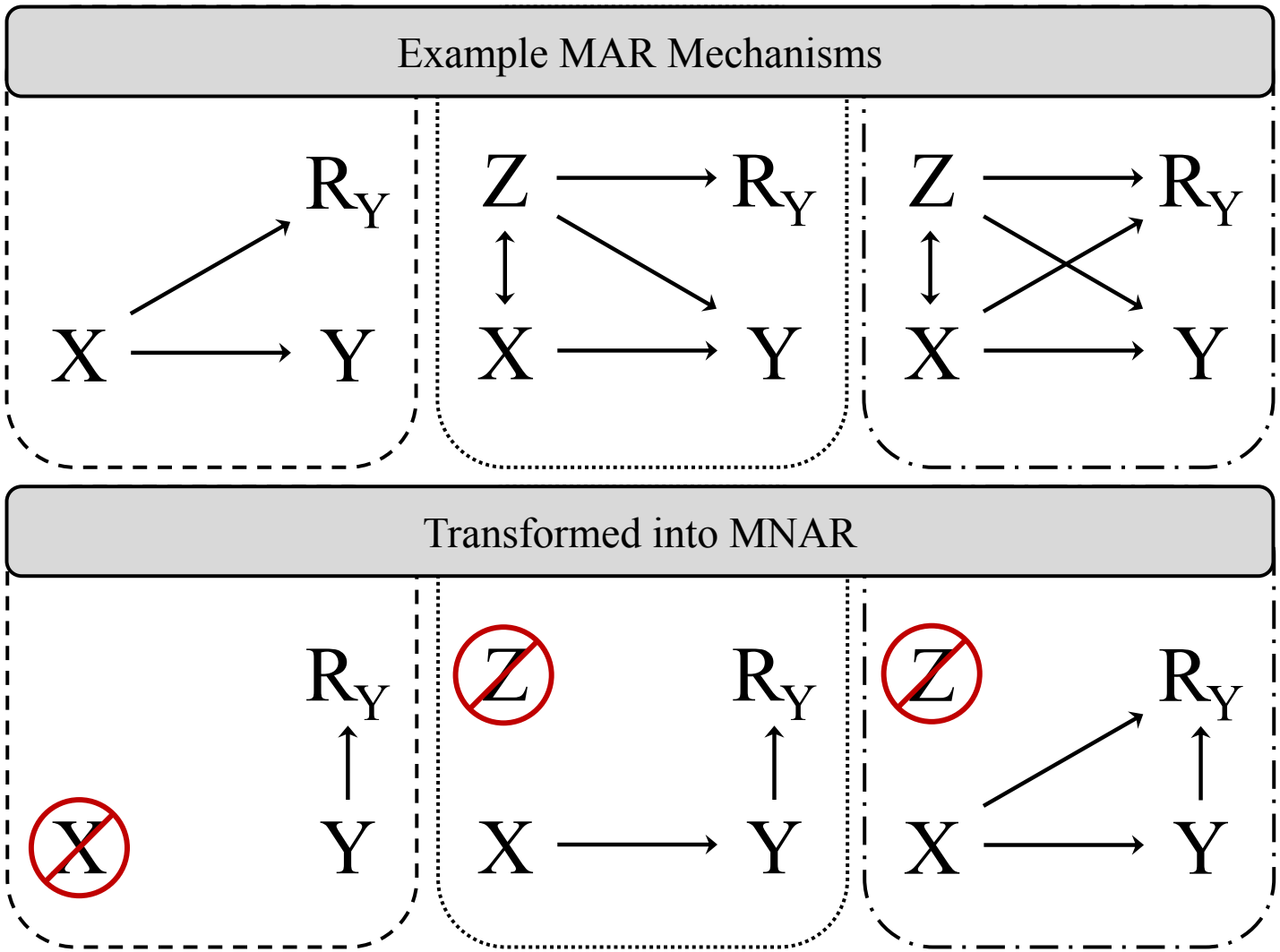
- A *Web of Science* search for citations of Von Hippel (2007) returns 297 results.
- Filtering those hits to only psychology related subjects results in 79 citations.
- Of these 79 papers, 60 (75.95%) employed the MID approach in empirical research.

Crux of the Matter

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

- The special case noted above amounts to a situation wherein the inferential regression model satisfies 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.

Graphical Representations



Study 1

MAR Missingness on Y

Methods: Simulation Parameters

Primary parameters

1. Proportion of the MAR predictor represented in the analysis model:
   - $pMAR = \{1.0, 0.75, 0.5, 0.25, 0.0\}$
2. Strength of correlations among the predictors:
   - $rXZ = \{0.0, 0.1, 0.3, 0.5\}$

Methods: Simulation Parameters

Primary parameters

1. Proportion of the MAR predictor represented in the analysis model:
   - $pMAR = \{1.0, 0.75, 0.5, 0.25, 0.0\}$
2. Strength of correlations among the predictors:
   - $rXZ = \{0.0, 0.1, 0.3, 0.5\}$

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: $R^2 = \{0.15, 0.3, 0.6\}$

Methods: Simulation Parameters

Primary parameters

1. Proportion of the MAR predictor represented in the analysis model:
   - $pMAR = \{1.0, 0.75, 0.5, 0.25, 0.0\}$
2. Strength of correlations among the predictors:
   - $rXZ = \{0.0, 0.1, 0.3, 0.5\}$

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.

Methods: 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.

Methods: 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: $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X + \hat{\beta}_2 Z_1$.

Methods: 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 a bivariate MAR predictor defined as:

$$\mathbf{Z}' = \gamma Z_1 + (1 - \gamma) Z_2, \ \gamma \in pMAR.$$

- Y values in the positive tail of $P(\mathbf{Z}')$ and X values in the negative tail of $P(\mathbf{Z}')$ were set to missing data.

Methods: Outcome Measures

The focal parameter was the slope coefficient for X (i.e., β_1).

Methods: Outcome Measures

The focal parameter was the slope coefficient for X (i.e., β_1).

For this report, I will focus on two outcome measures:

1. Percentage Relative Bias:

$$PRB = 100\left(\frac{\bar{\hat{\beta}}_1 - \beta_1}{\beta_1}\right)$$

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.

Methods: Computational Details

The simulation code was written in the R statistical programming language (R Core Team, 2017).

Methods: Computational Details

The simulation code was written in the R statistical programming language (R Core Team, 2017).

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

- $M = 100$ imputations
- $I = 10$ iterations of the MICE algorithm
- Bayesian linear regressions as elementary imputation methods

Methods: Computational Details

The simulation code was written in the R statistical programming language (R Core Team, 2017).

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

- $M = 100$ imputations
- $I = 10$ iterations of the MICE algorithm
- Bayesian linear regressions as elementary imputation methods

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

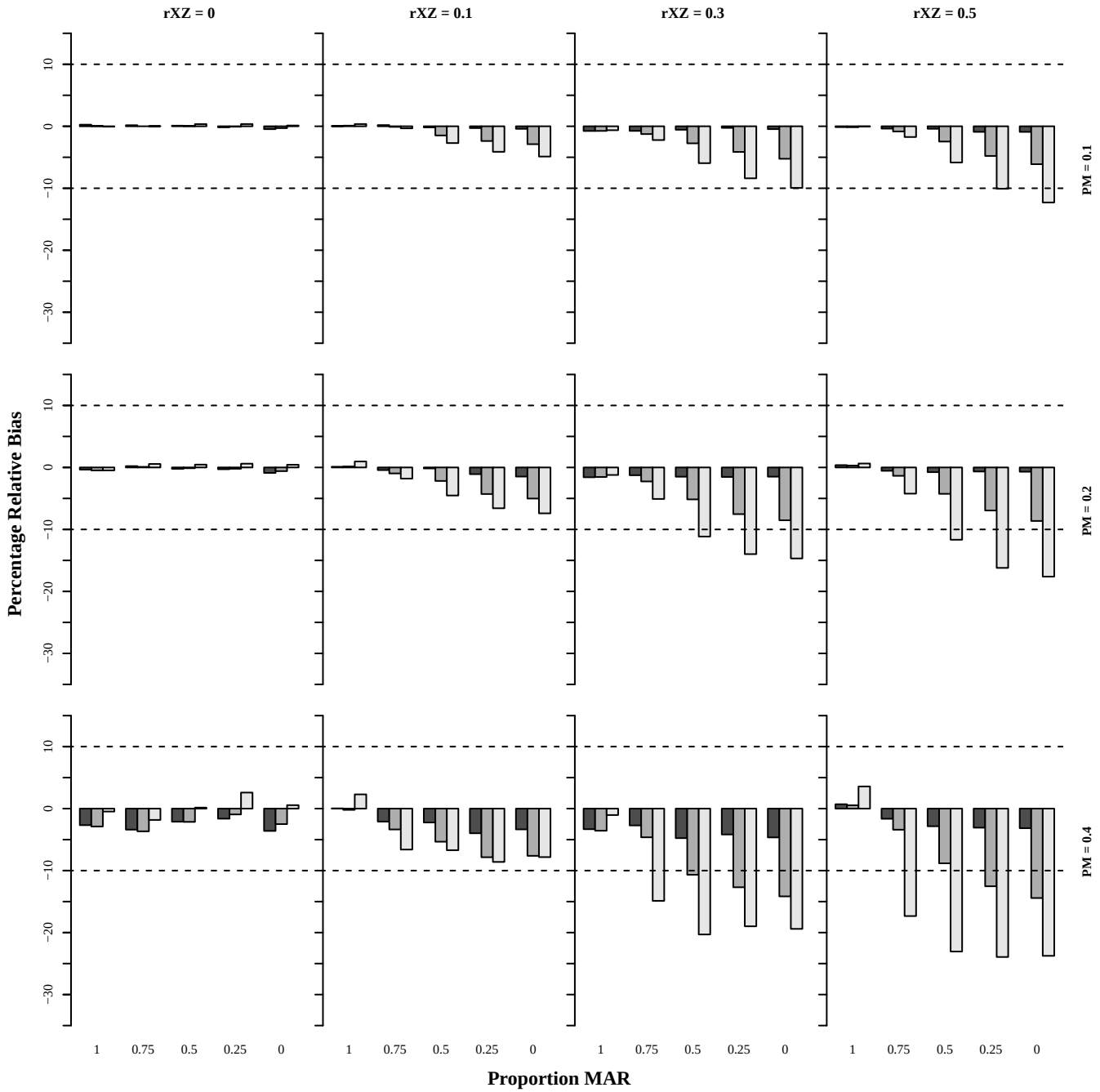
Methods: Hypotheses

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.

Study 1 Results

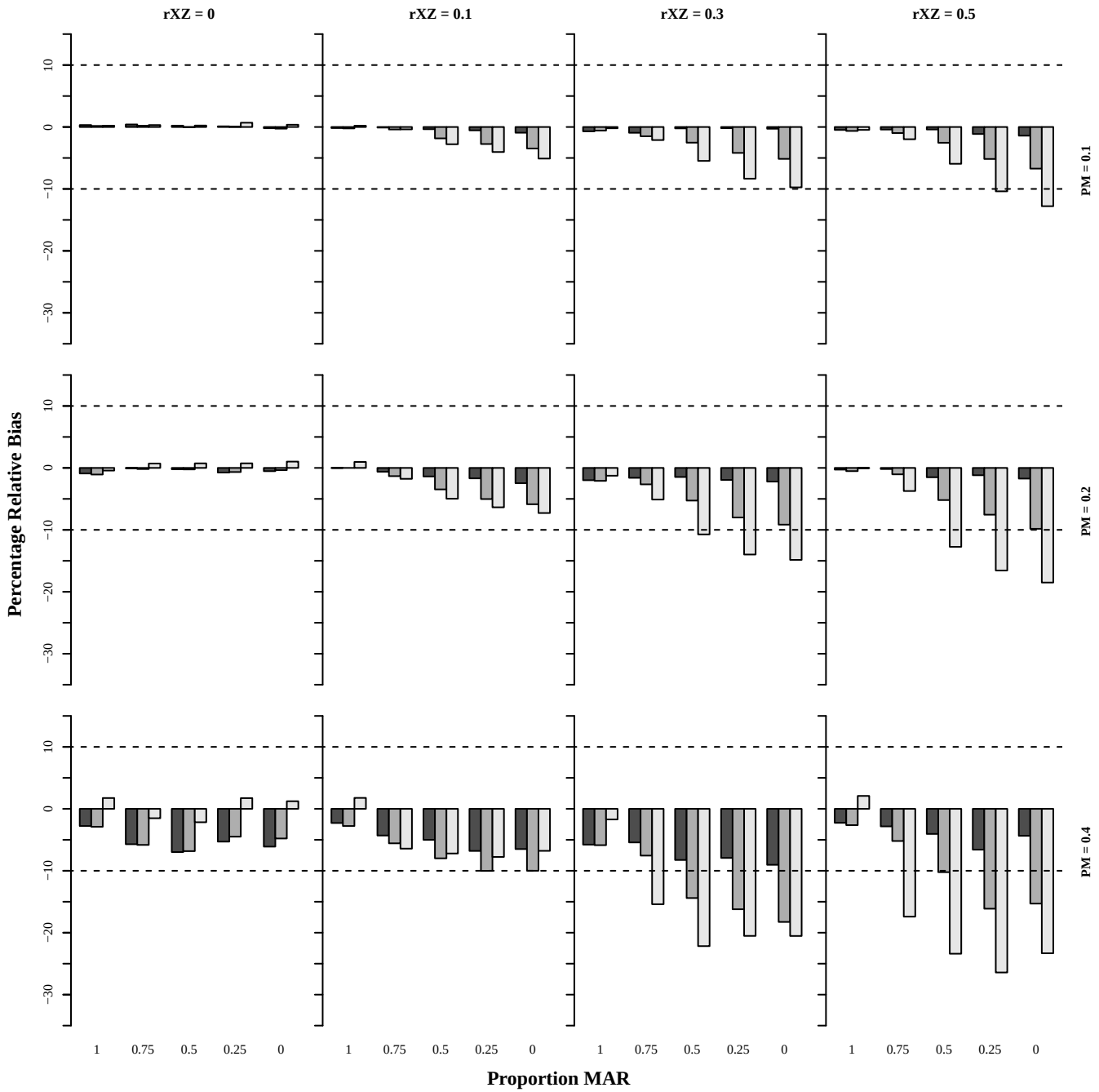
Percentage Relative Bias

- $N = 500$
- $R^2 = 0.3$



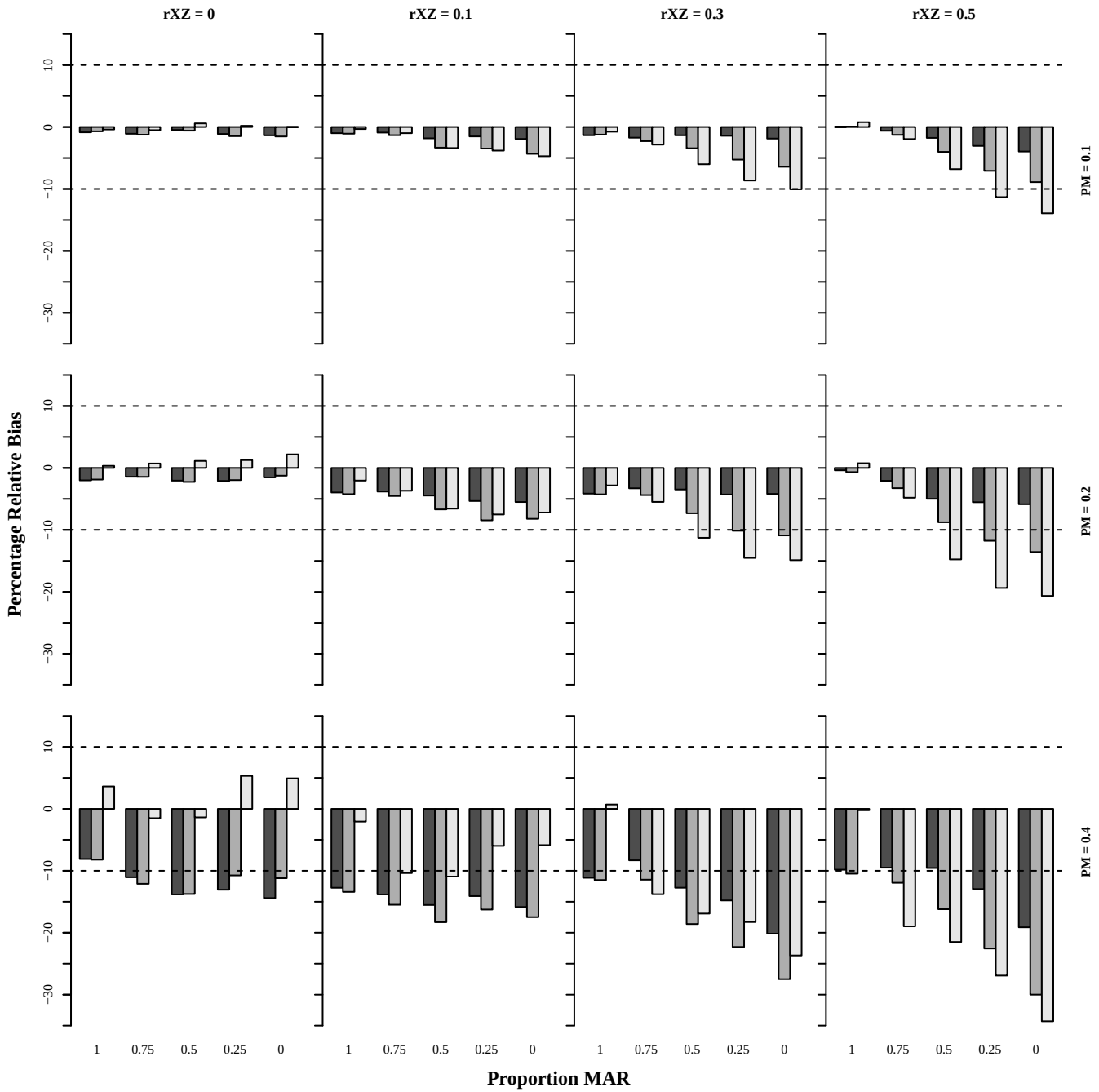
Percentage Relative Bias

- $N = 250$
- $R^2 = 0.3$



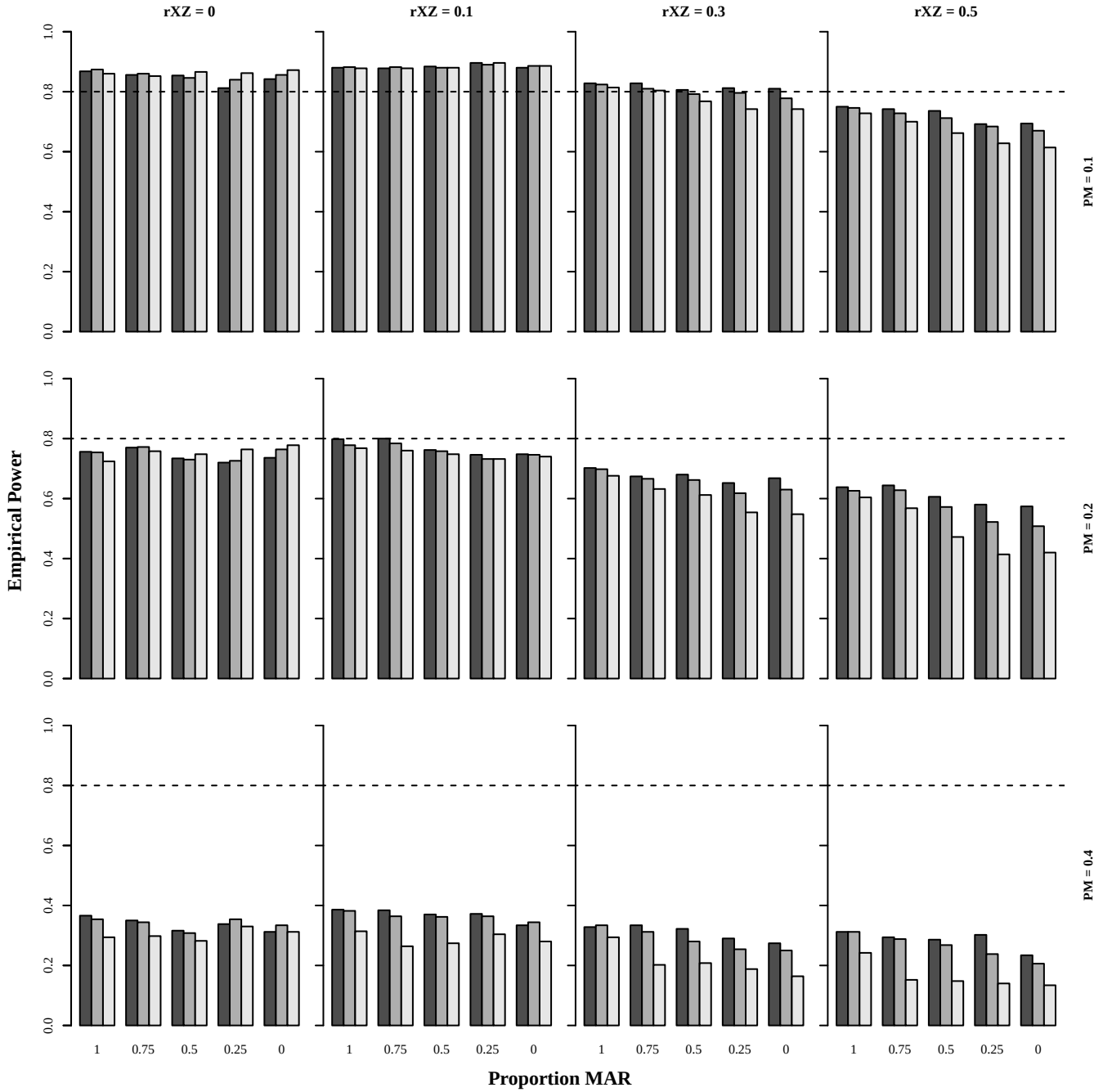
Percentage Relative Bias

- $N = 100$
- $R^2 = 0.3$



Empirical 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 performed badly*.
  - *LWD produced minimal bias with MCAR data and when $\mathbf{Z}' = Z_1$

Discussion: Bias-Related Findings

Traditional MI did not lead to bias in most conditions.

- When $N = 100$ and $PM = 0.4$ all methods performed badly*.
  - *LWD produced minimal bias with MCAR data and when $\mathbf{Z}' = Z_1$

Whenever $\mathbf{Z}' \neq Z_1$, MID and LWD produced biased estimates of regression slopes.

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

Discussion: Bias-Related Findings

Traditional MI did not lead to bias in most conditions.

- When $N = 100$ and $PM = 0.4$ all methods performed badly*.
  - *LWD produced minimal bias with MCAR data and when $\mathbf{Z}' = Z_1$

Whenever $\mathbf{Z}' \neq Z_1$, MID and LWD produced biased estimates of regression slopes.

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

Scientific preference for parsimonious models $\longrightarrow$ MAR predictors frequently excluded from analytic IVs

Discussion: Power-Related Findings

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

- Taking M large enough 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).

Discussion: Power-Related Findings

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

- Taking M large enough 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).

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

- Both MID and LWD entail throwing away data.

Study 2

MNAR Missingness on Y

Methods: Simulation Parameterization

The Study 2 simulation parameters, data-generating models, and analysis model, were the same as those in Study 1.

- Only the types of missing data differed.

Methods: Simulation Parameterization

The Study 2 simulation parameters, data-generating models, and analysis model, were the same as those in Study 1.

- Only the types of missing data differed.
  - MAR missingness was imposed on X using the same form of MAR predictor employed in Study 1:

$$\mathbf{Z}' = \gamma Z_1 + (1 - \gamma) Z_2, \ \gamma \in pMAR$$

Methods: Simulation Parameterization

The Study 2 simulation parameters, data-generating models, and analysis model, were the same as those in Study 1.

- Only the types of missing data differed.
  - MAR missingness was imposed on X using the same form of MAR predictor employed in Study 1:

$$\mathbf{Z}' = \gamma Z_1 + (1 - \gamma) Z_2, \ \gamma \in pMAR$$

  - MNAR missingness was imposed on Y as a function of Y itself.
    - Y values in the positive tail of $P(\mathbf{Z}')$ and X values in the negative tail of $P(\mathbf{Z}')$ were set to missing data.
    - The *pMAR* parameter only affected missingness on X.

Methods: Hypotheses

1. All three methods will produce biased estimates of regression coefficients.

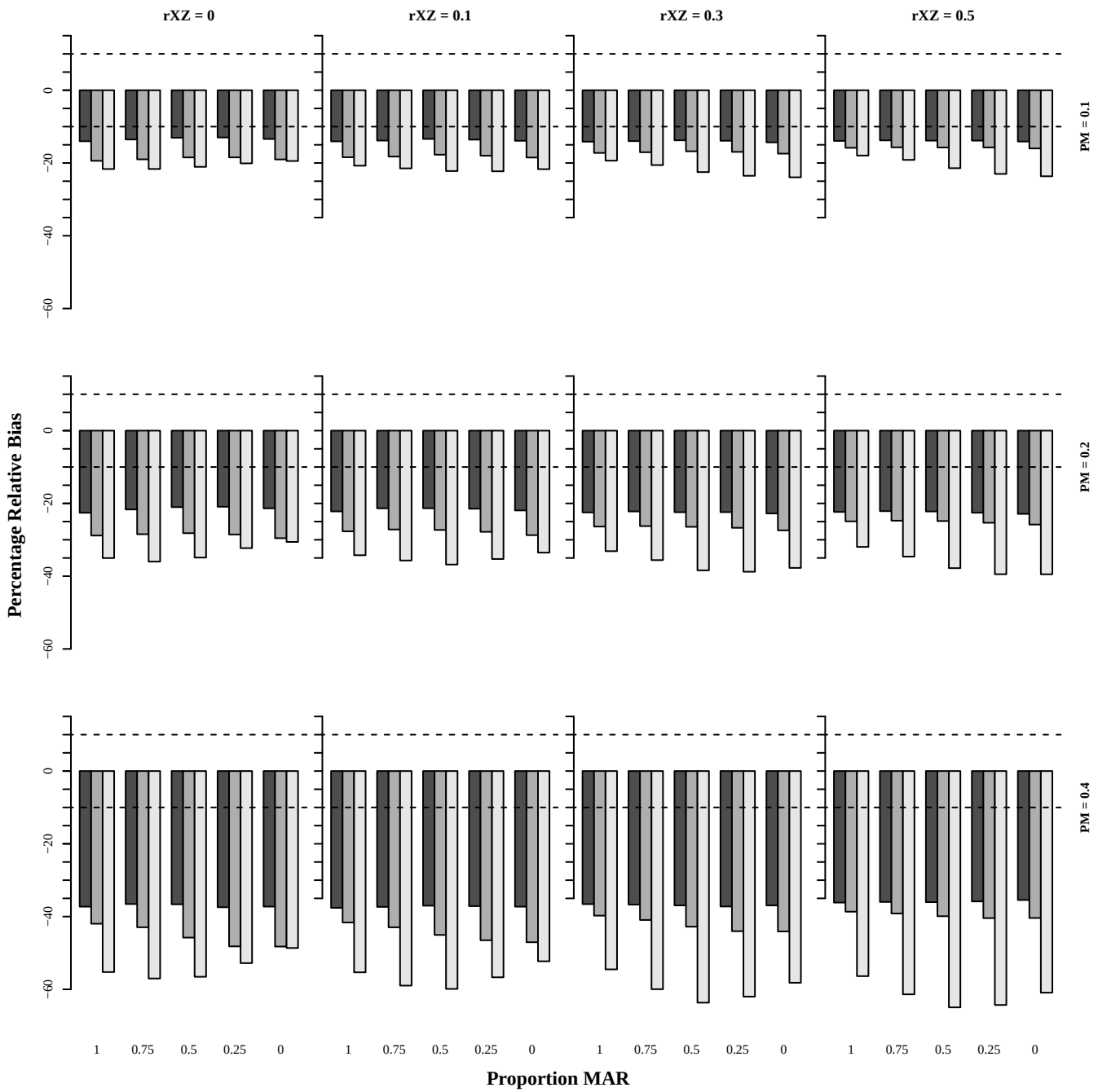
2. The bias produced by traditional MI will be no worse than the bias produced by MID and LWD.

3. Traditional MI will maintain power levels at least as high as the power levels produced by MID and LWD.

Study 2 Results

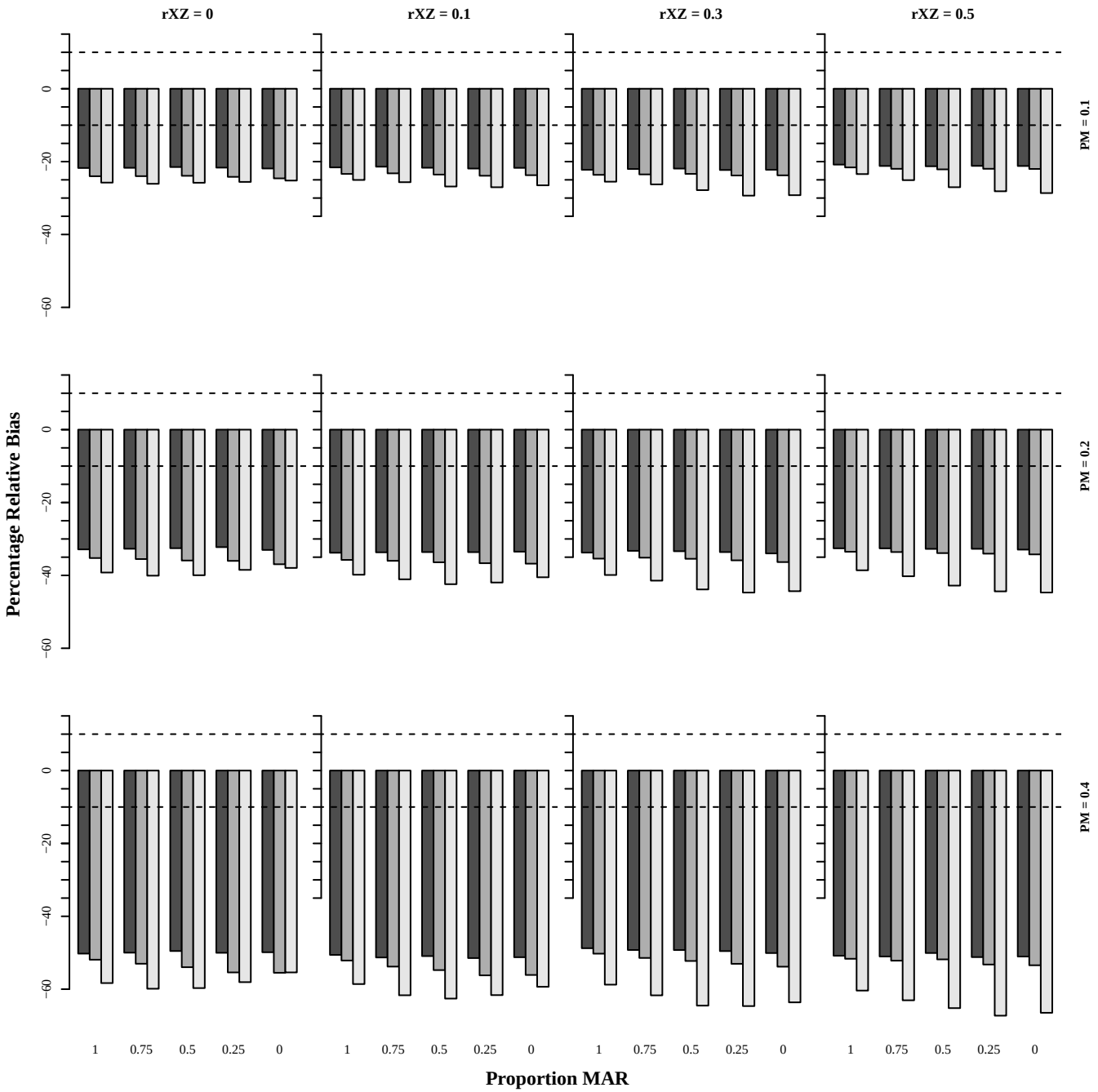
Percentage Relative Bias

- $N = 100$
- $R^2 = 0.6$



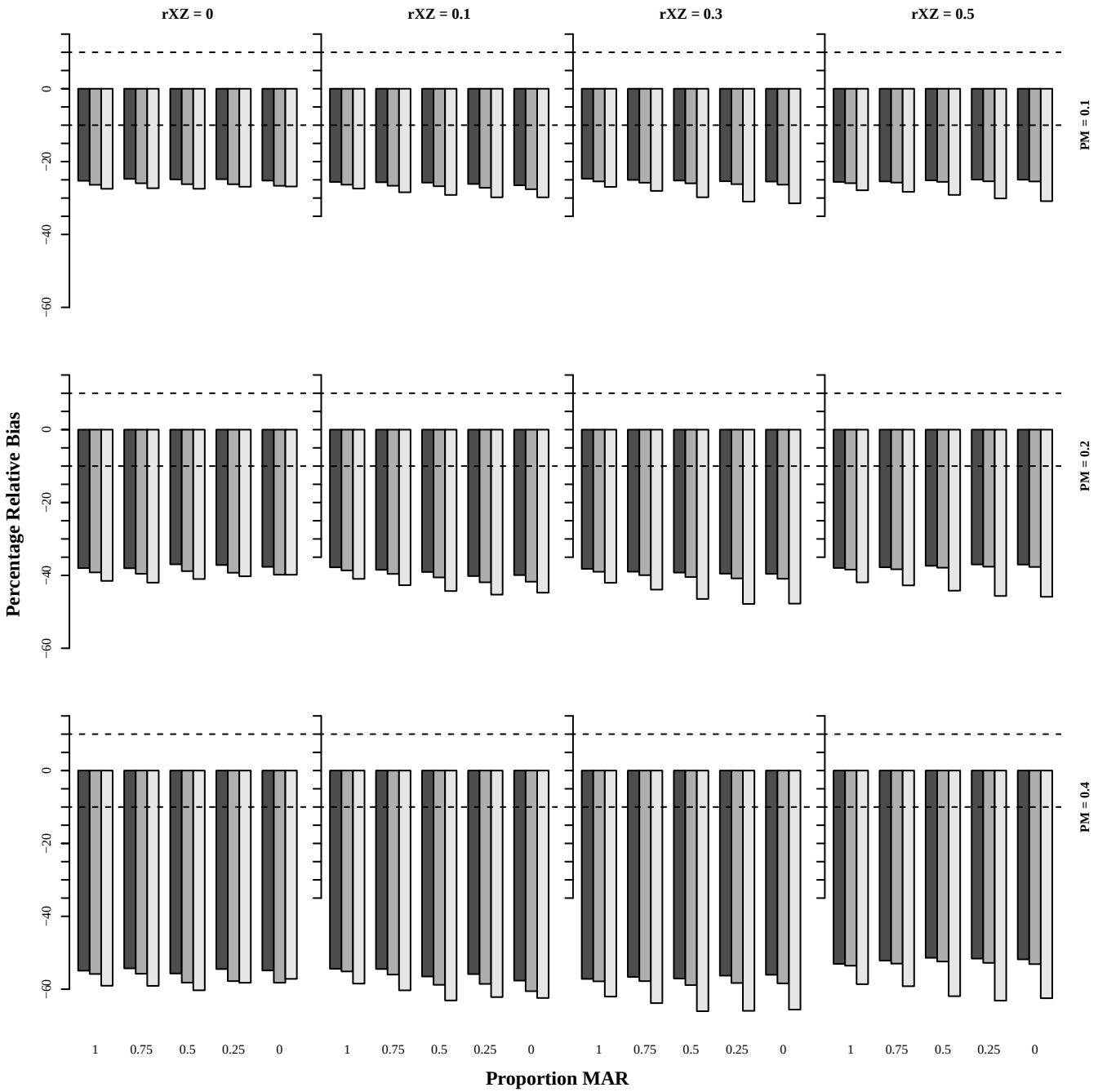
Percentage Relative Bias

- $N = 100$
- $R^2 = 0.3$



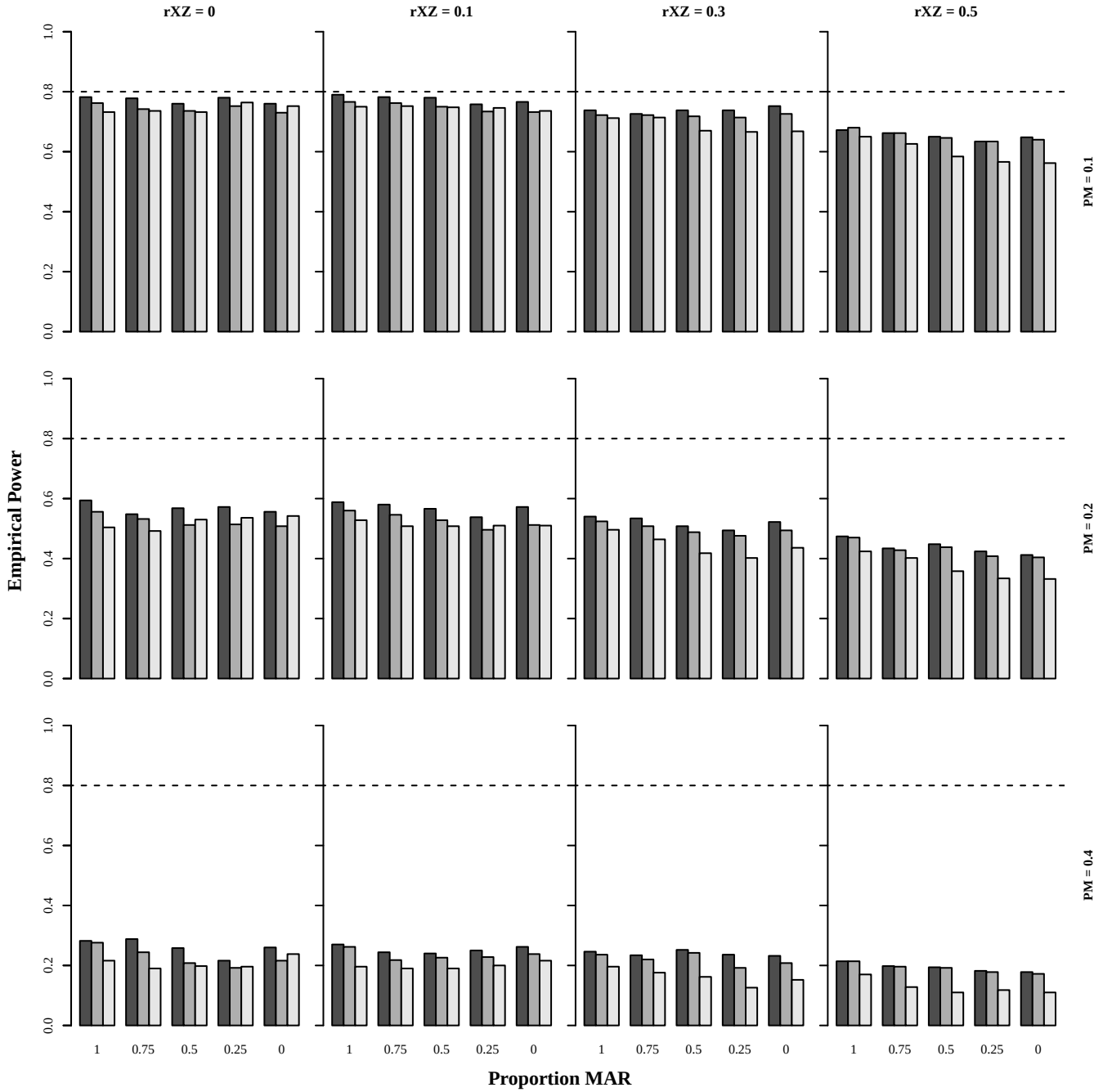
Percentage Relative Bias

- $N = 100$
- $R^2 = 0.15$



Empirical Power

- $N = 100$
- $R^2 = 0.3$



Study 2 Discussion

All methods produced biased parameter estimates with MNAR missingness on Y.

- The biases induced by traditional MI were equal to or smaller than those induced by MID and LWD.

Power loss was no greater with traditional MI than it was with MID.

- Both traditional MI and MID produced better power than LWD.

General Conclusions

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.

General Conclusions

The methodological argument against traditional MI, in favor of MID, assumes 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.
  - Even with MNAR data, traditional MI tended to outperform MID and never produced greater biases.

Limitations

The models I employed were very simple.

- Some may question the ecological validity of these results.
- I have purposefully focused on internal validity.

Limitations

The models I employed were very simple.

- Some may question the ecological validity of these results.
- I have 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).

Take Home Message

Don't be Fancy. Impute your DVs!
(and don't delete observations, 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. (2017). 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.

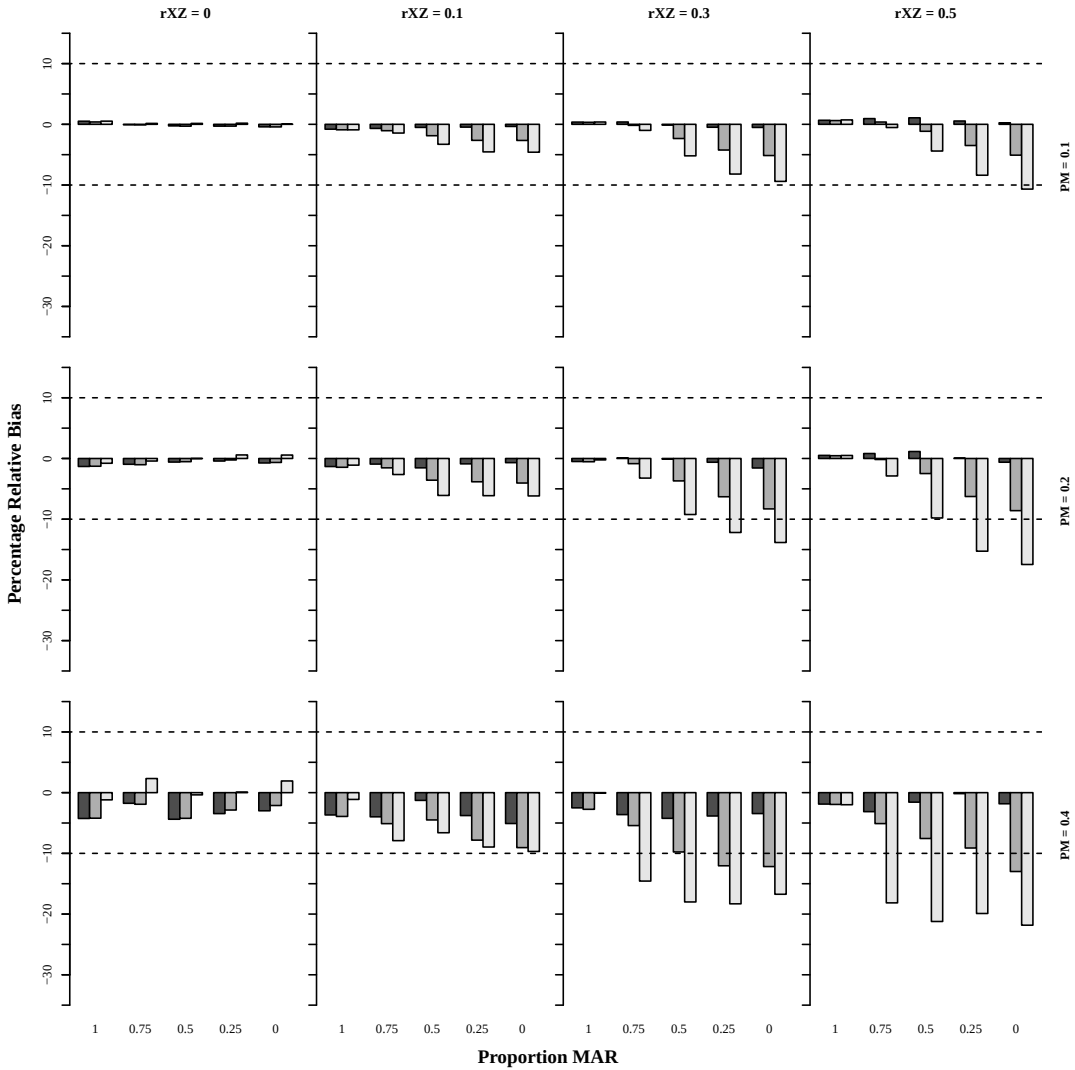
References

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

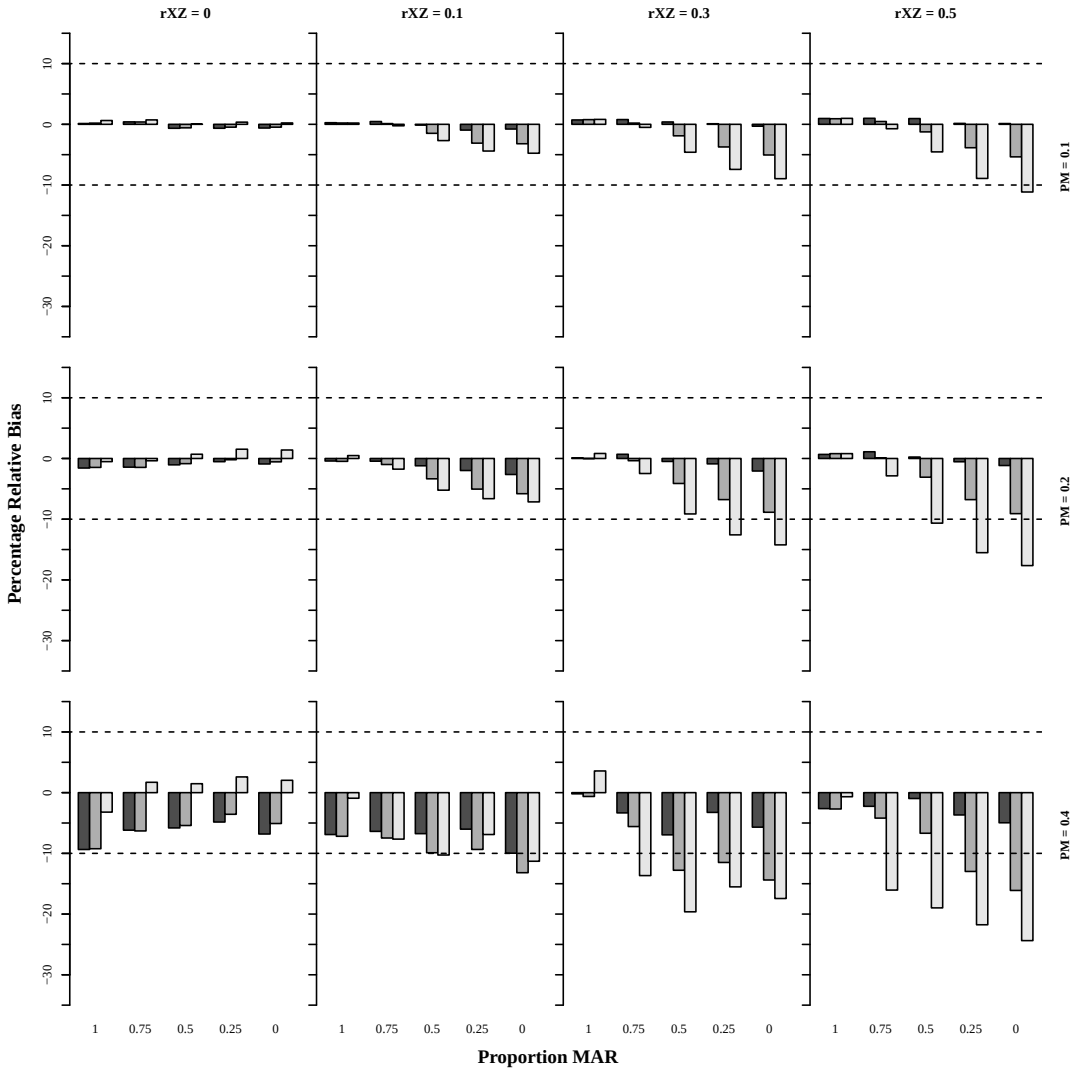
Von Hippel, P. T. (2007). Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, *37*(1), 83–117.

Extras

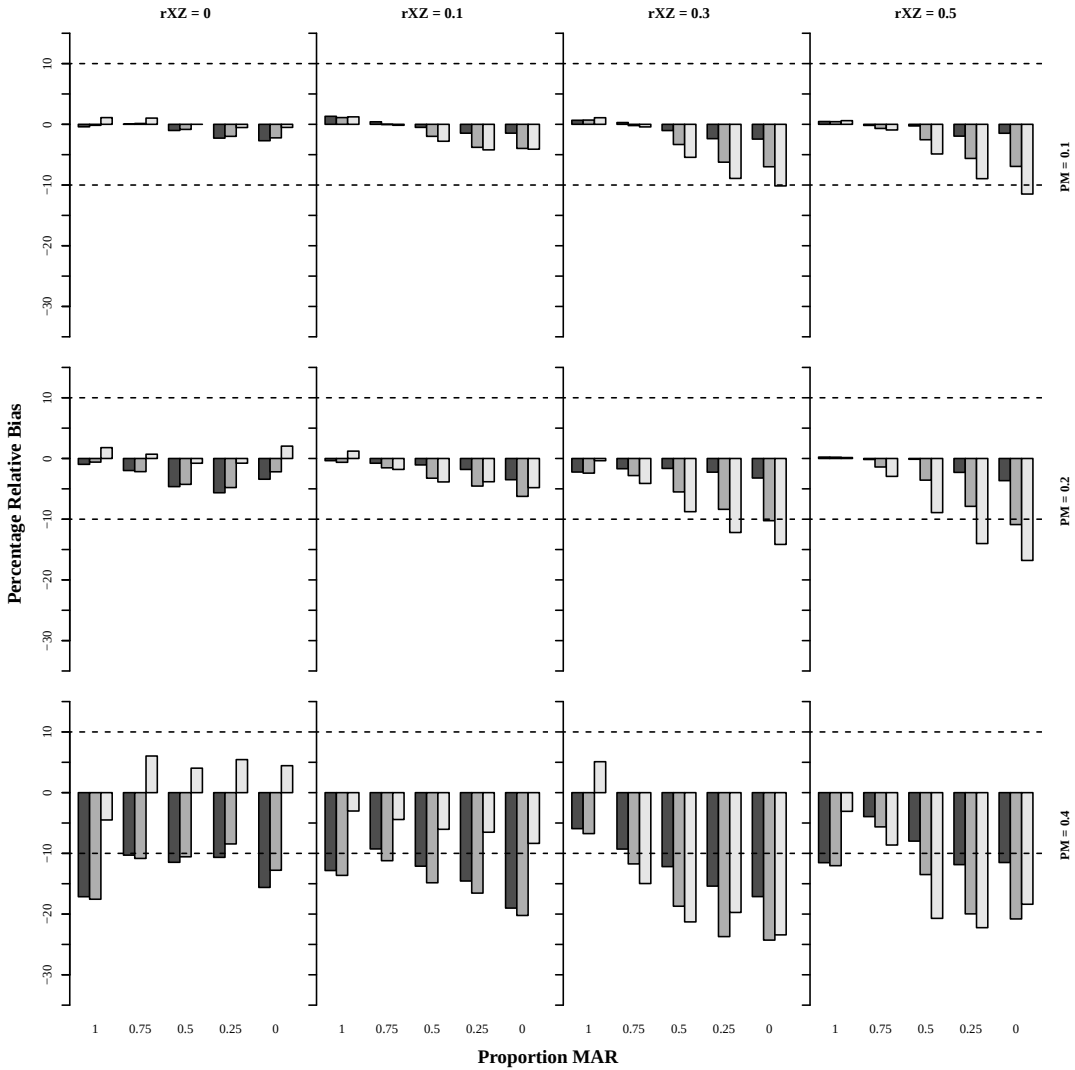
Study 1 PRB with $N = 500$, $R^2 = 0.15$



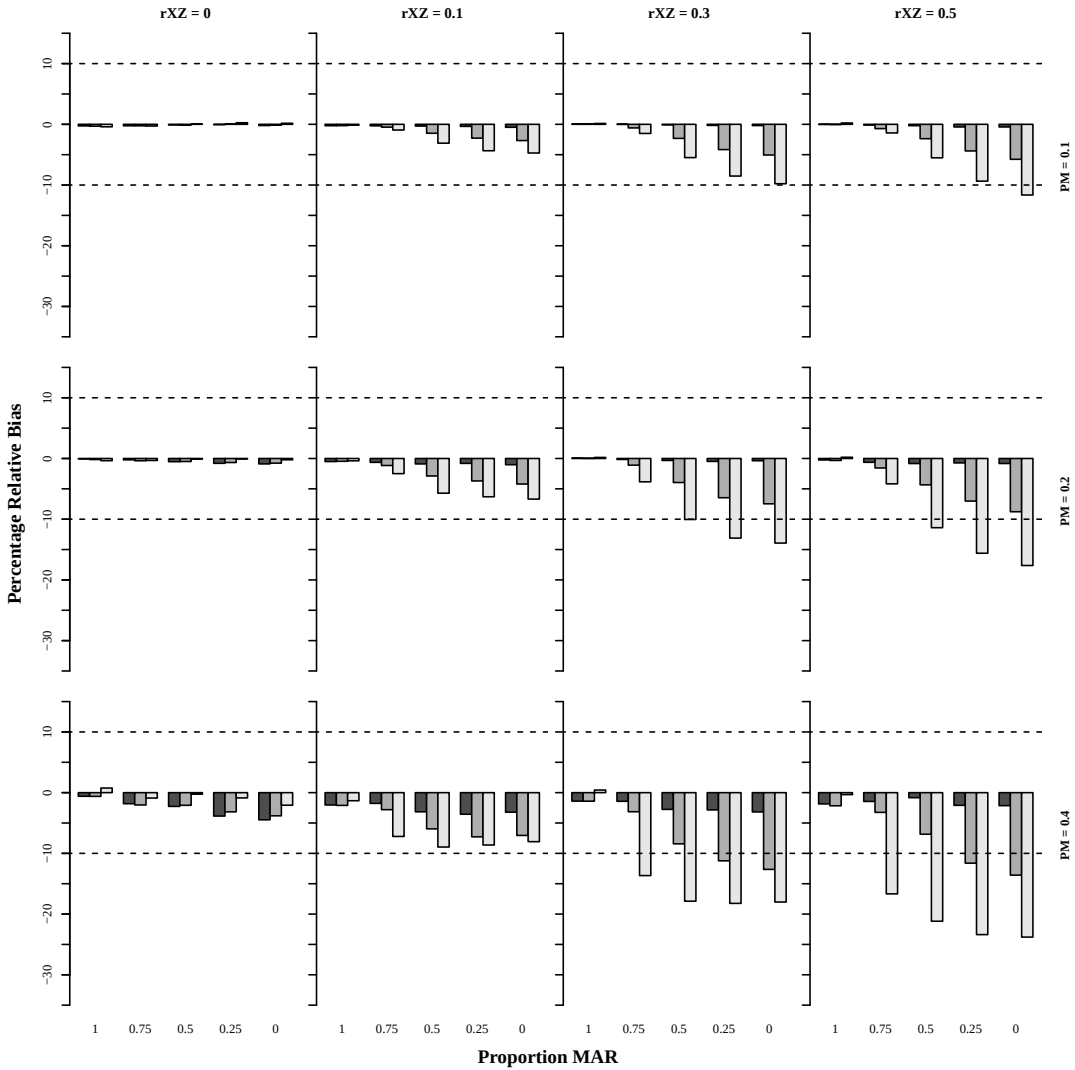
Study 1 PRB with $N = 250$, $R^2 = 0.15$



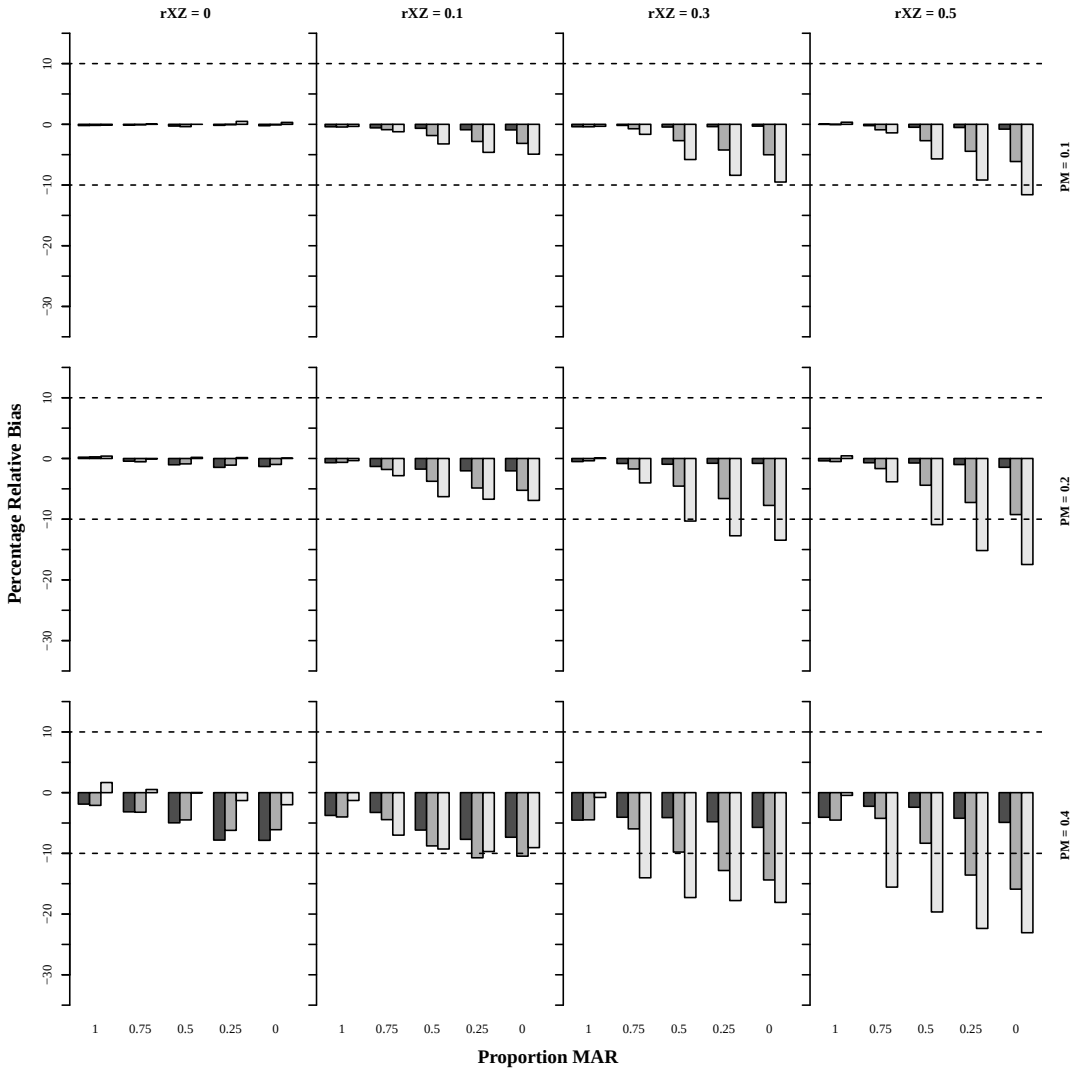
Study 1 PRB with $N = 100$, $R^2 = 0.15$



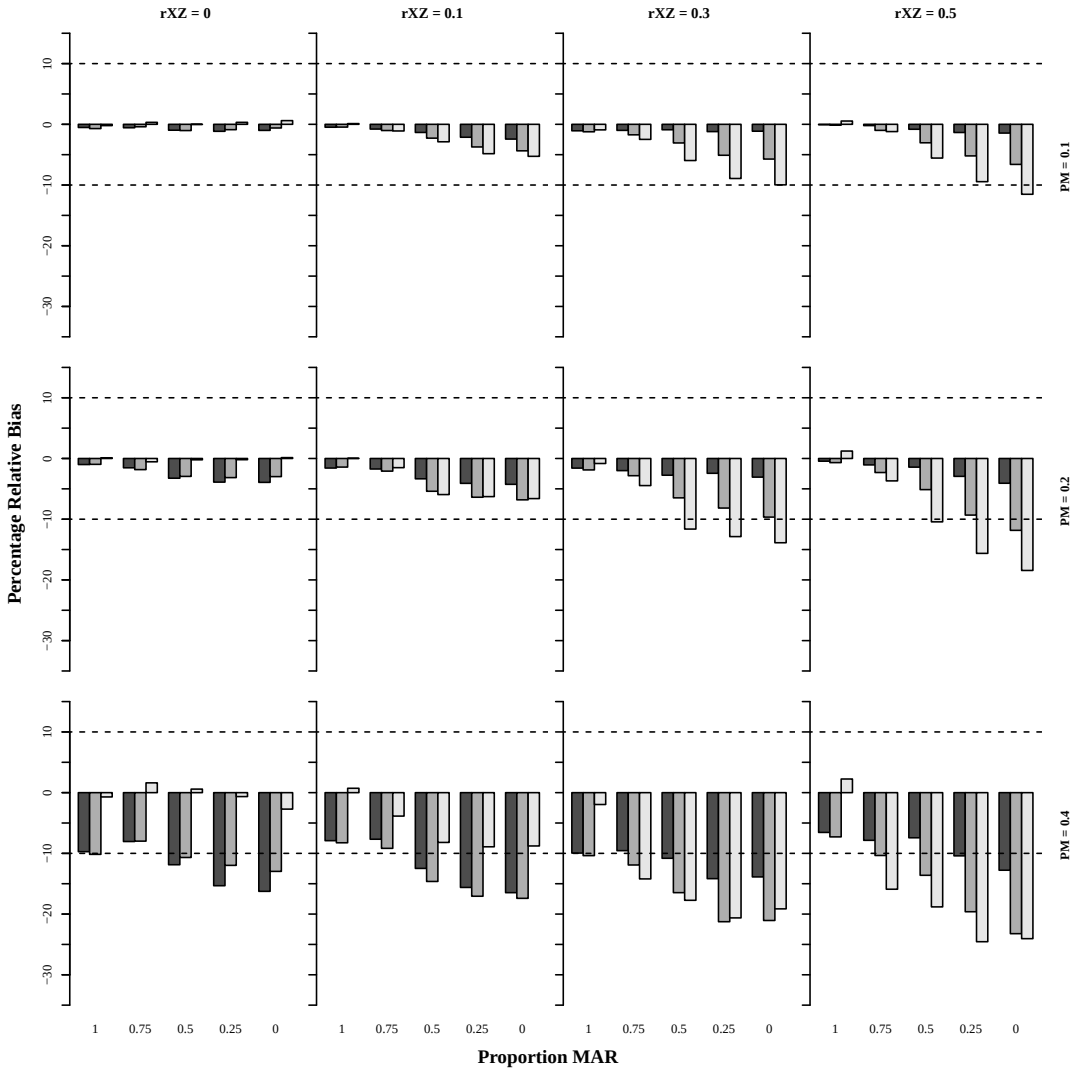
Study 1 PRB with $N = 500$, $R^2 = 0.60$



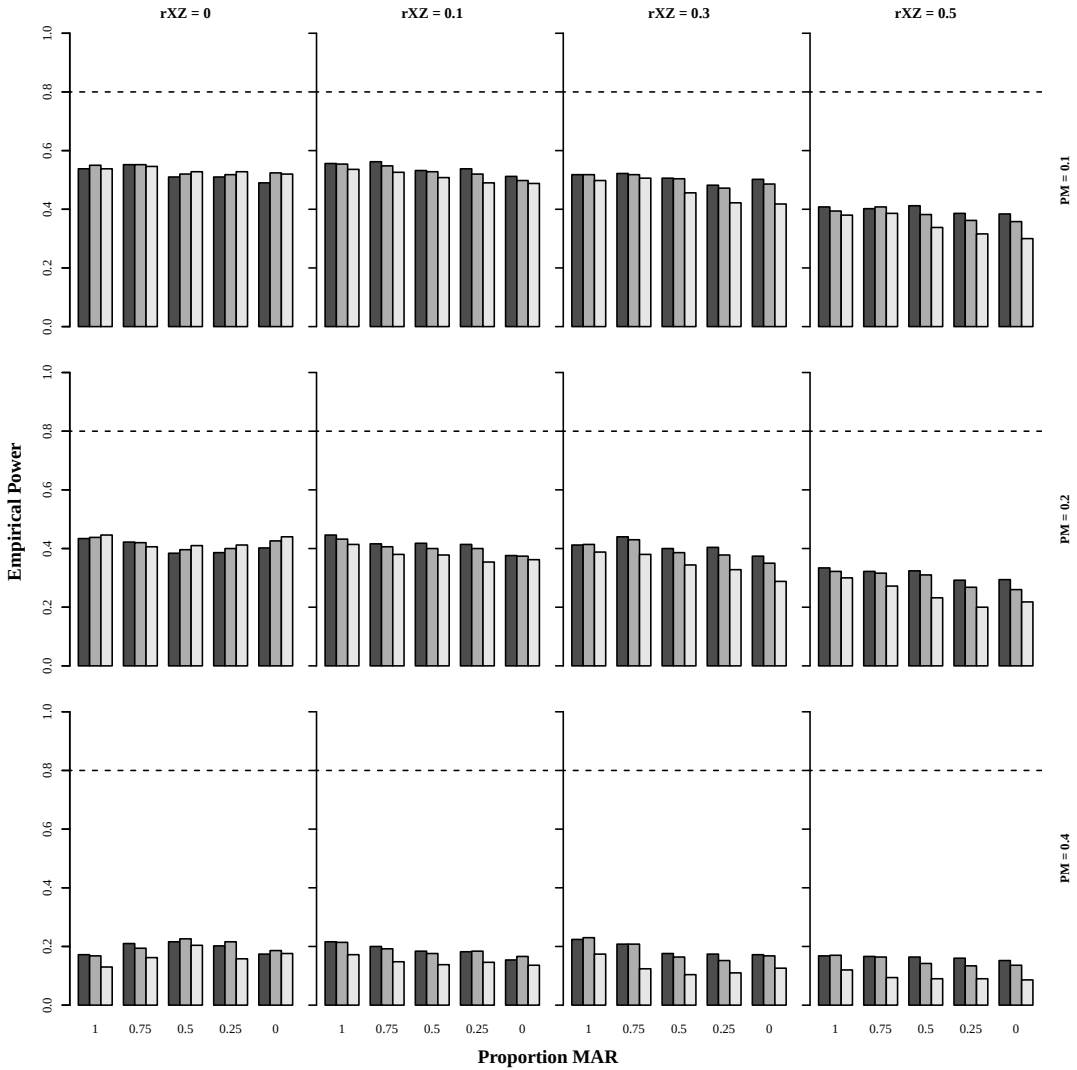
Study 1 PRB with $N = 250$, $R^2 = 0.60$



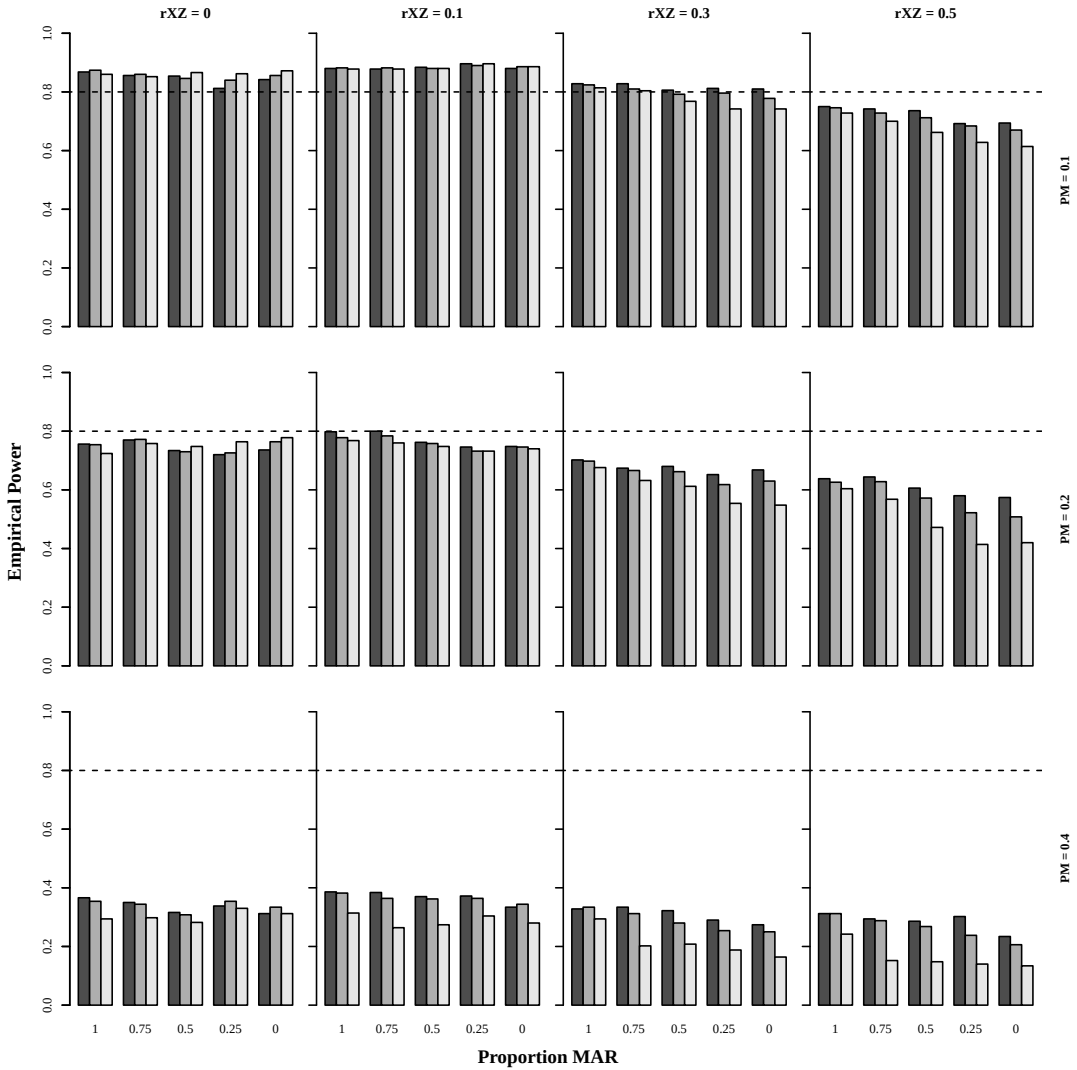
Study 1 PRB with $N = 100$, $R^2 = 0.60$



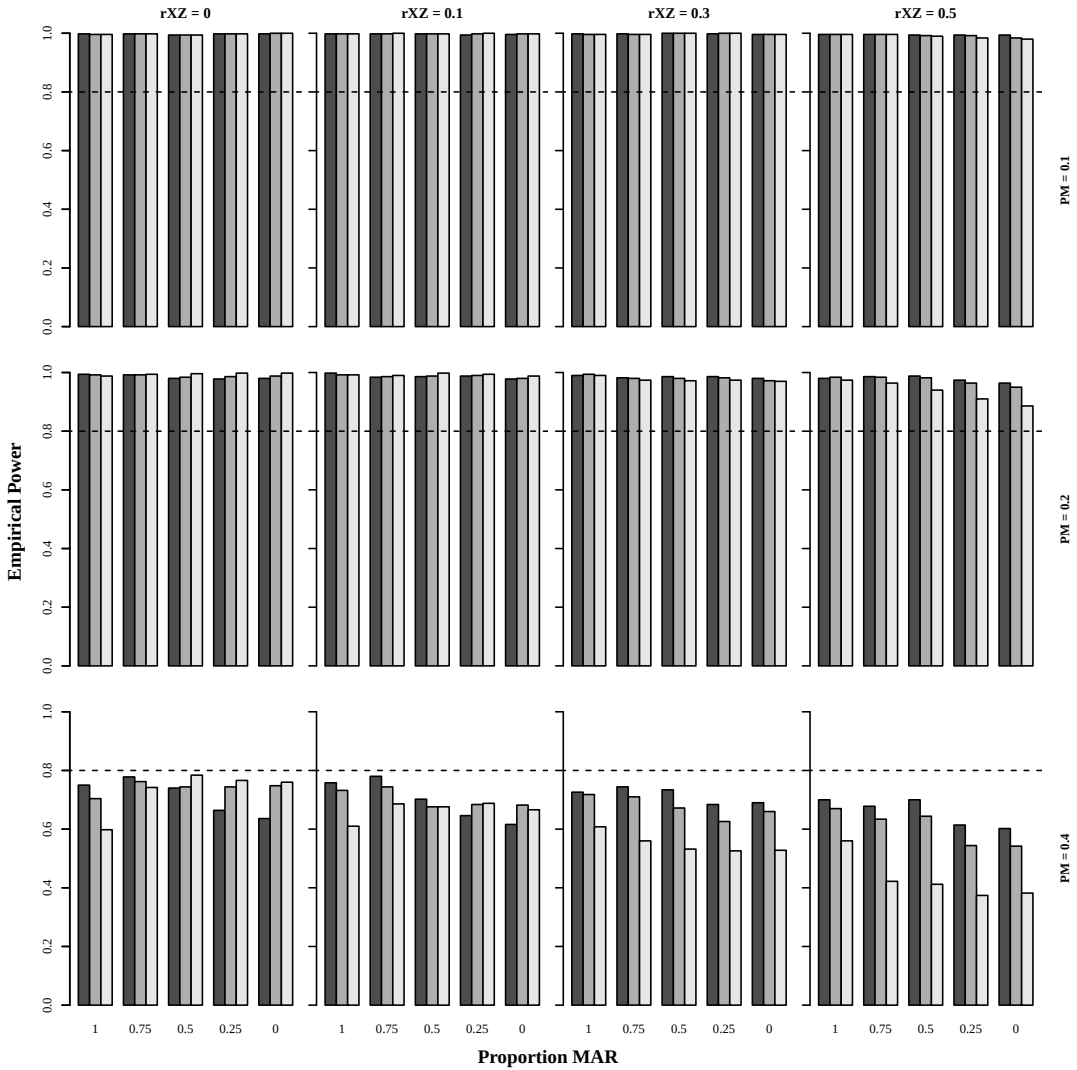
Study 1 Power with $N = 100$, $R^2 = 0.15$



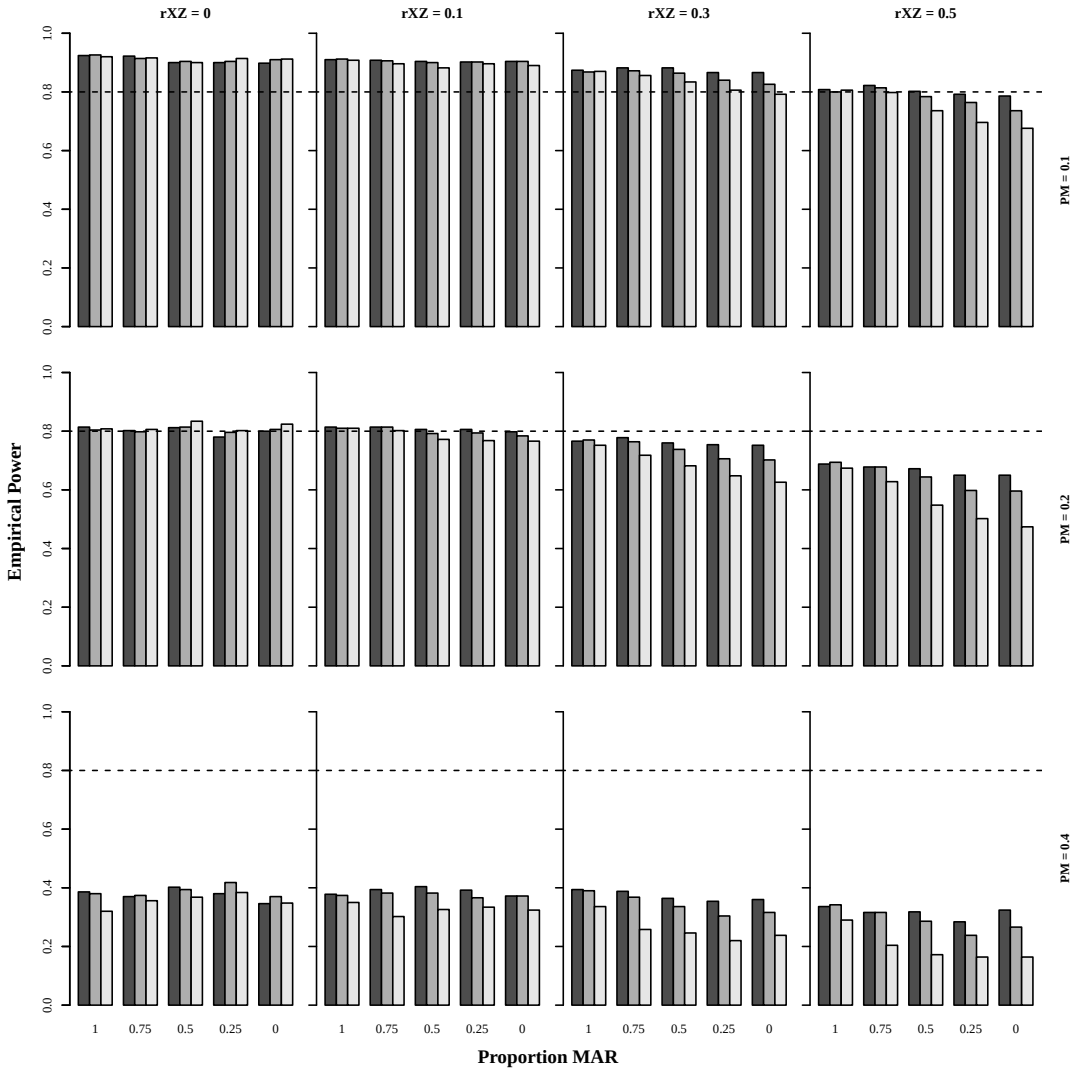
Study 1 Power with $N = 100$, $R^2 = 0.3$



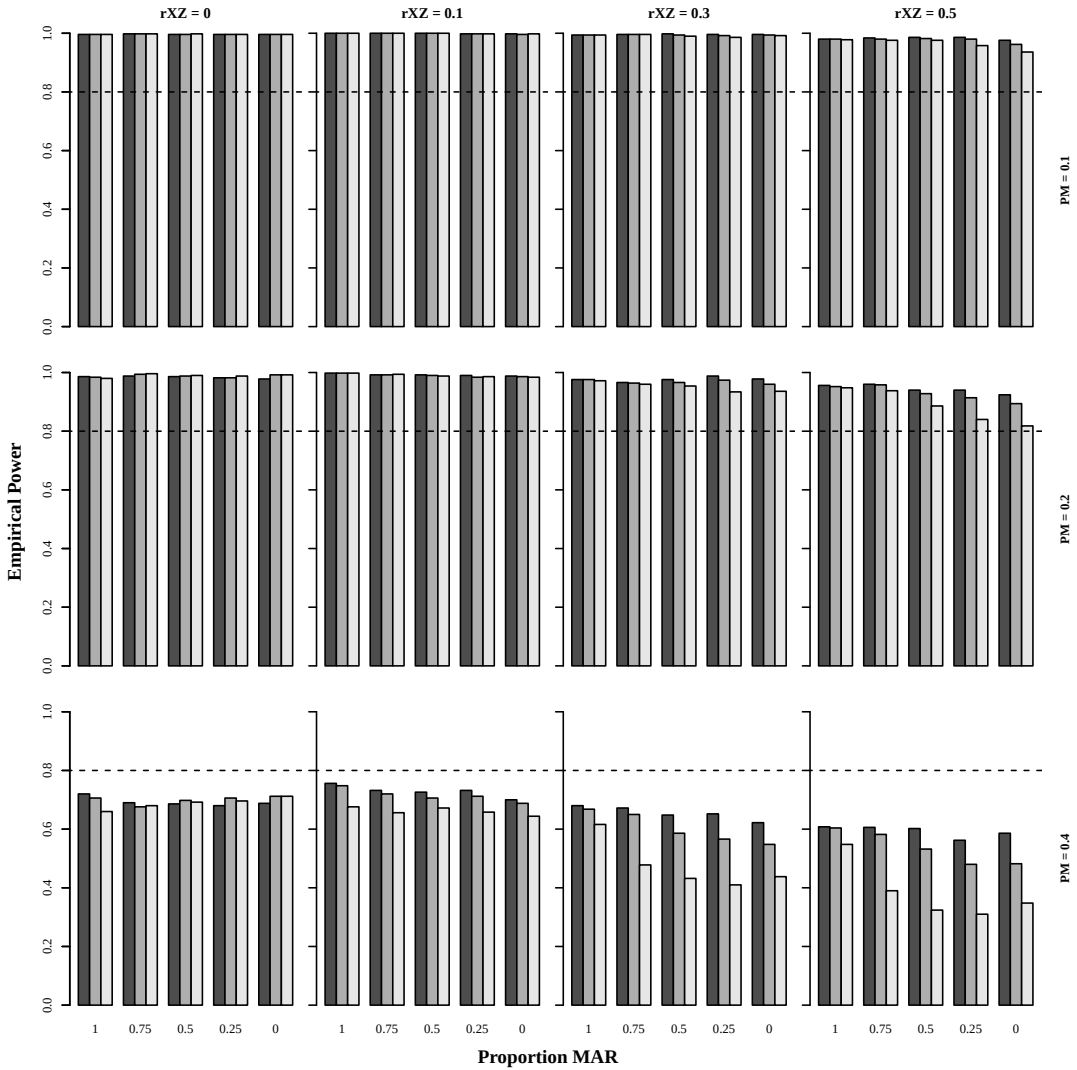
Study 1 Power with $N = 100$, $R^2 = 0.6$



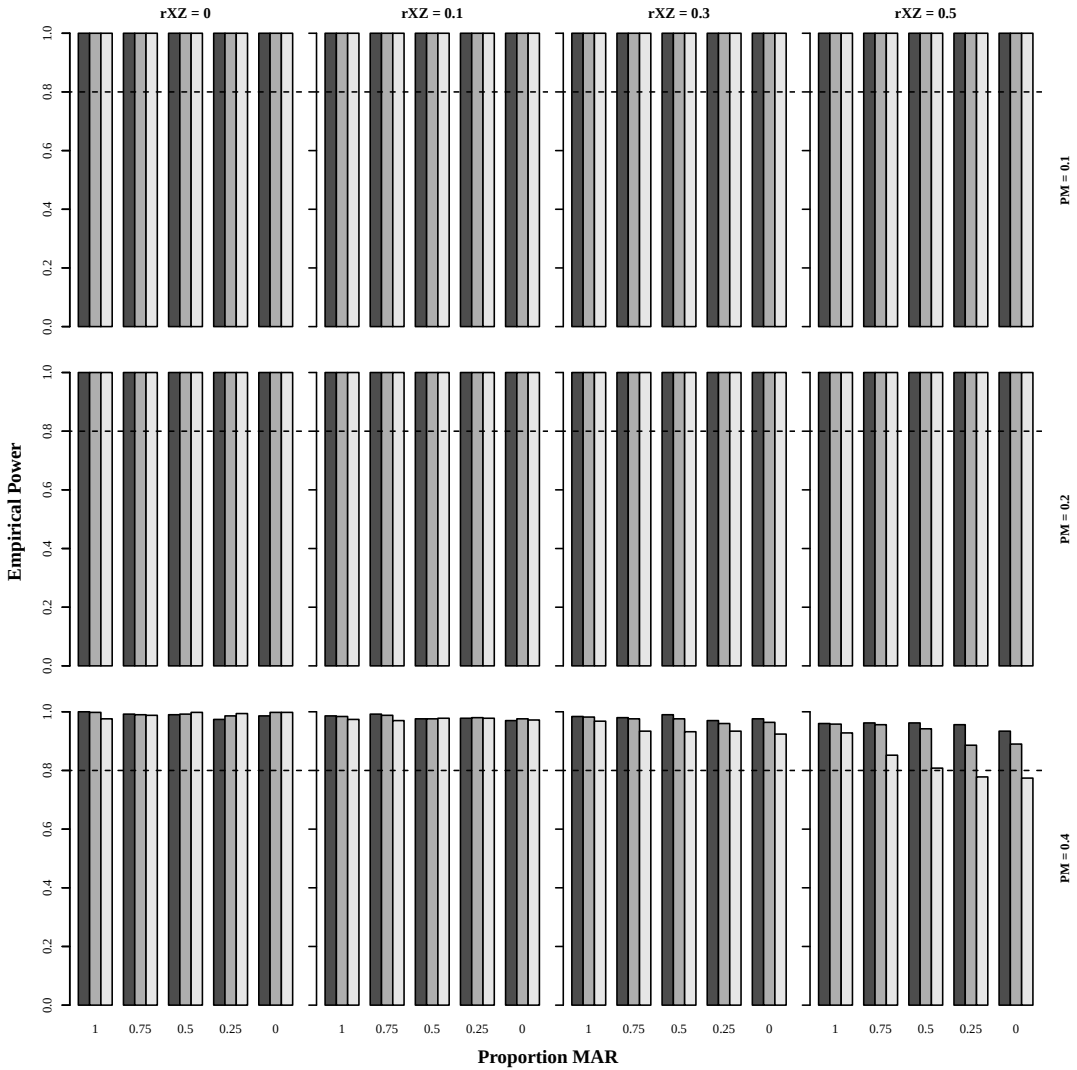
Study 1 Power with $N = 250$, $R^2 = 0.15$



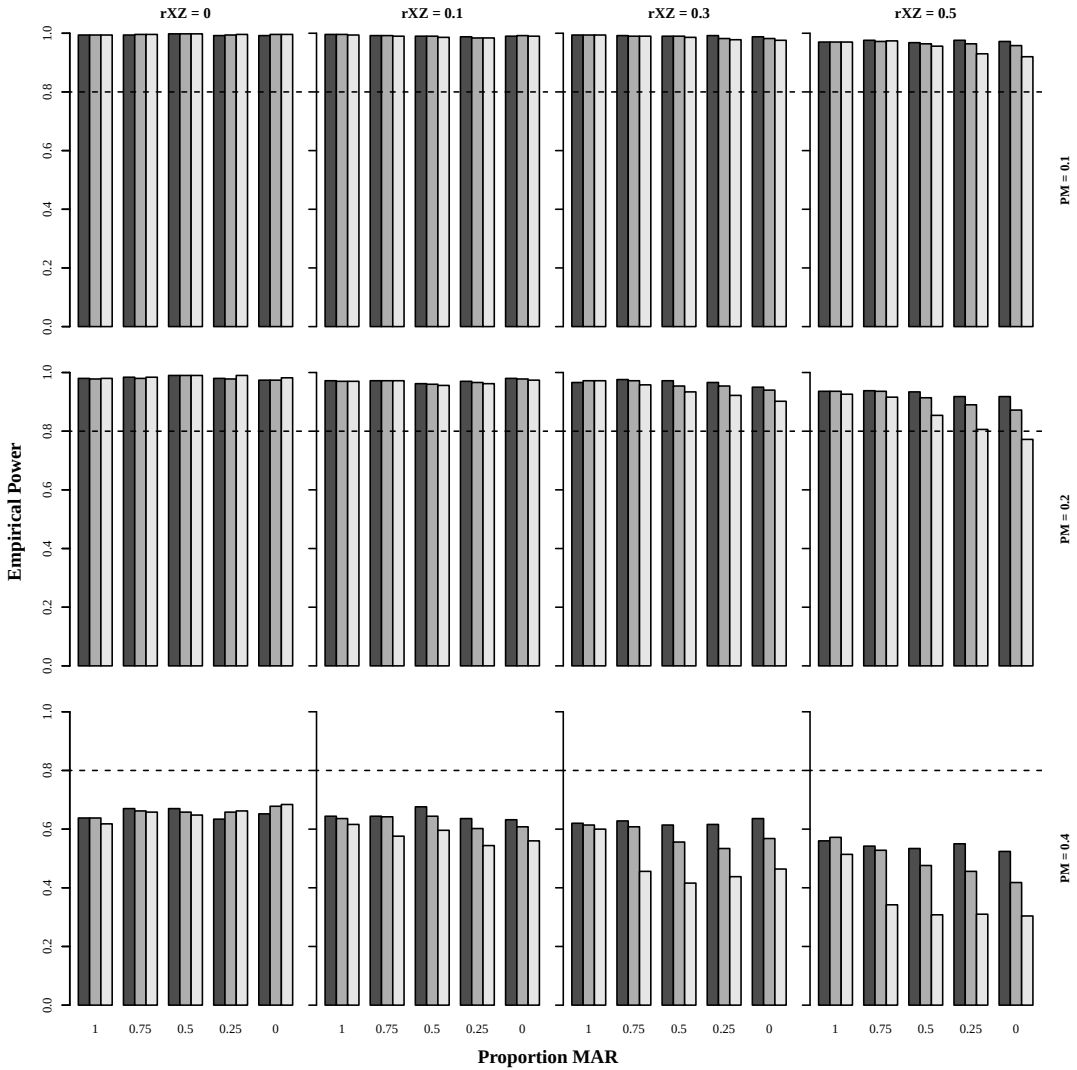
Study 1 Power with $N = 250$, $R^2 = 0.30$



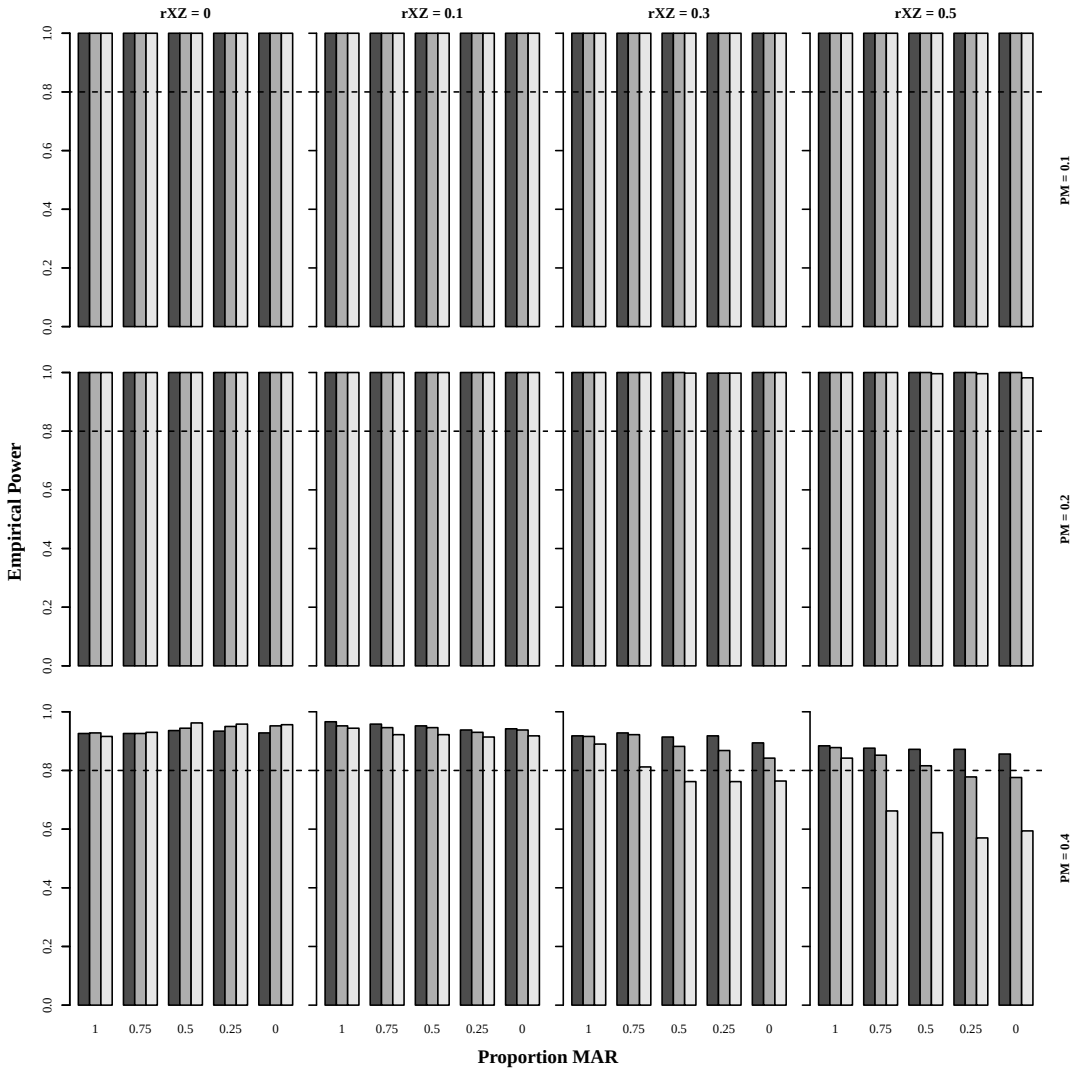
Study 1 Power with $N = 250$, $R^2 = 0.60$



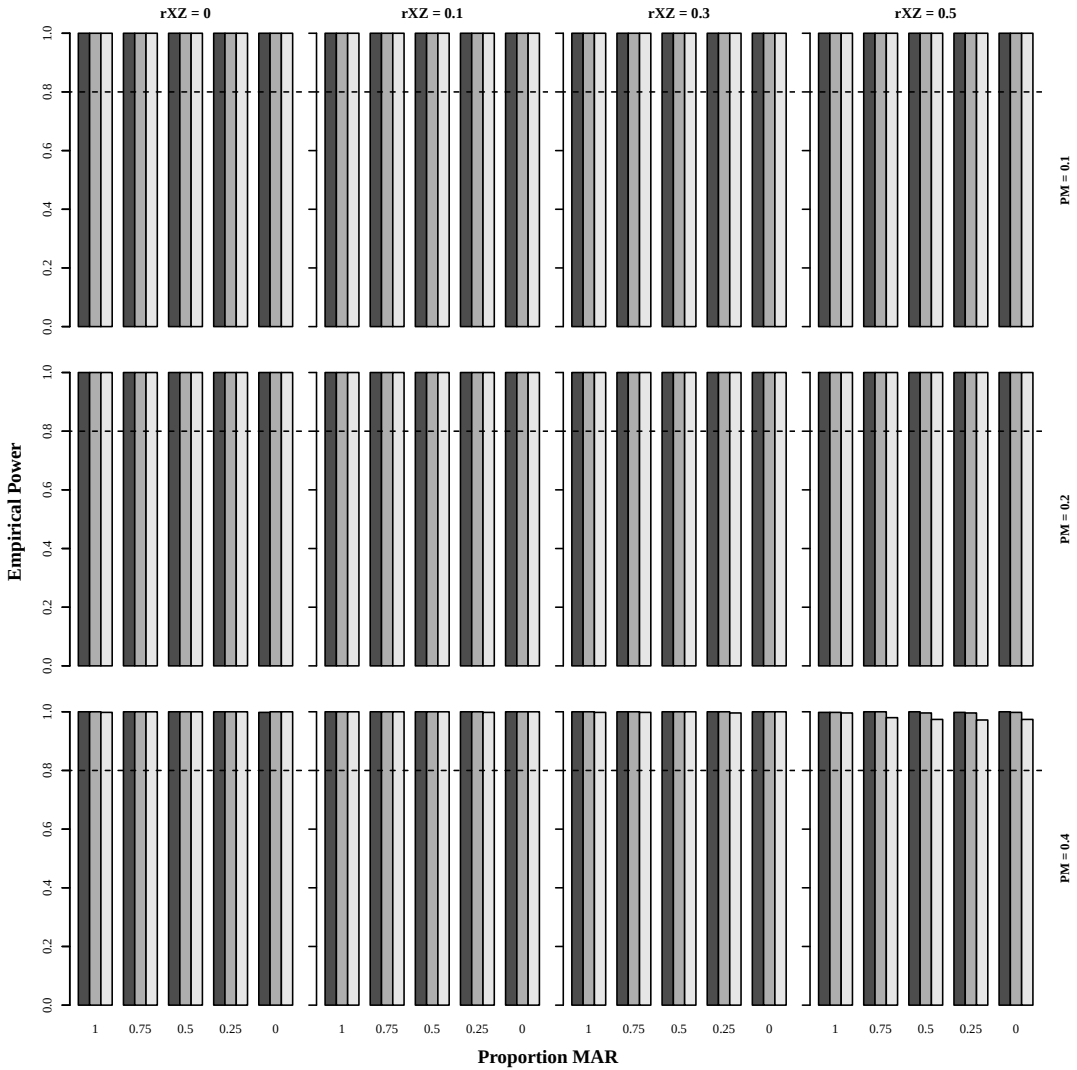
Study 1 Power with $N = 500$, $R^2 = 0.15$



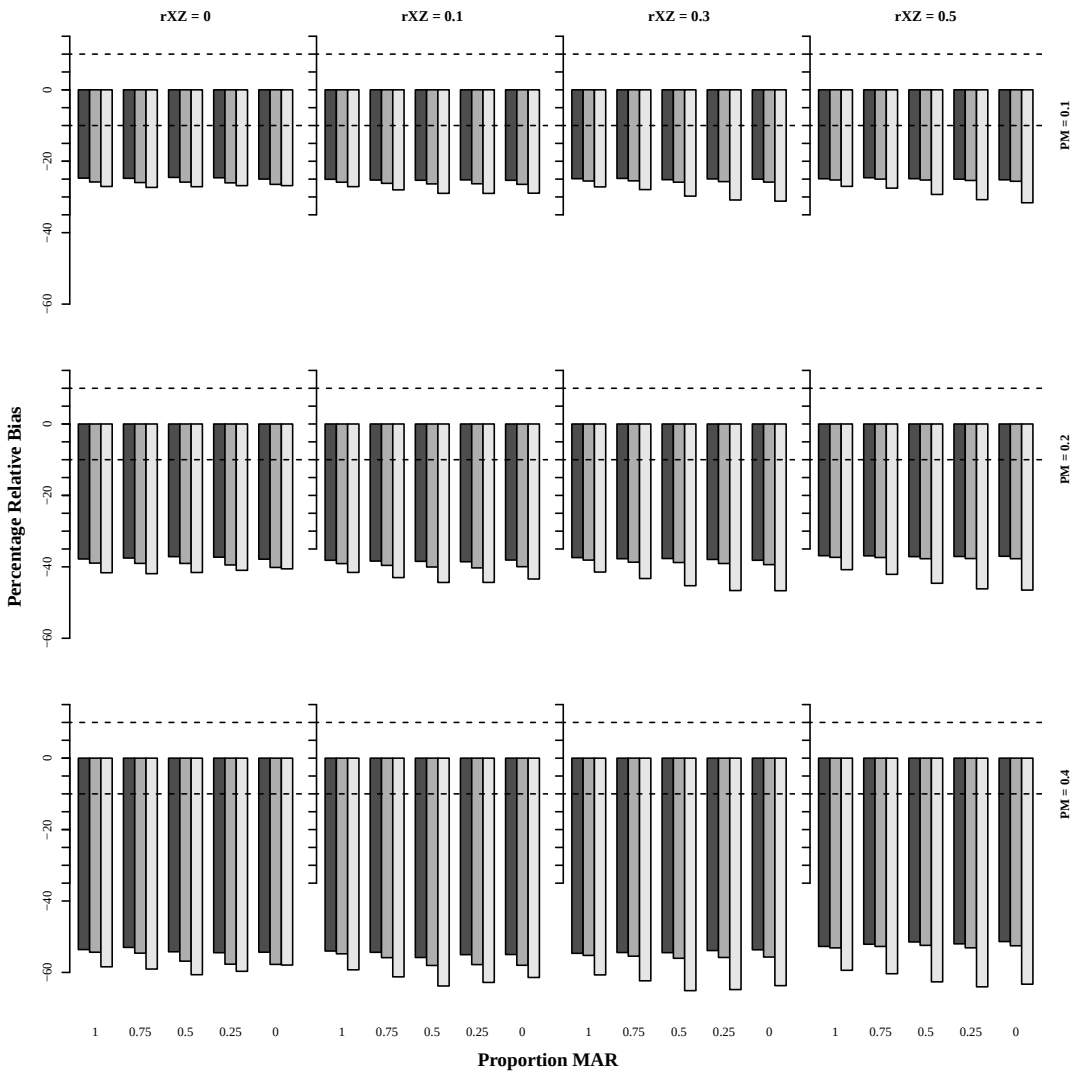
Study 1 Power with $N = 500$, $R^2 = 0.30$



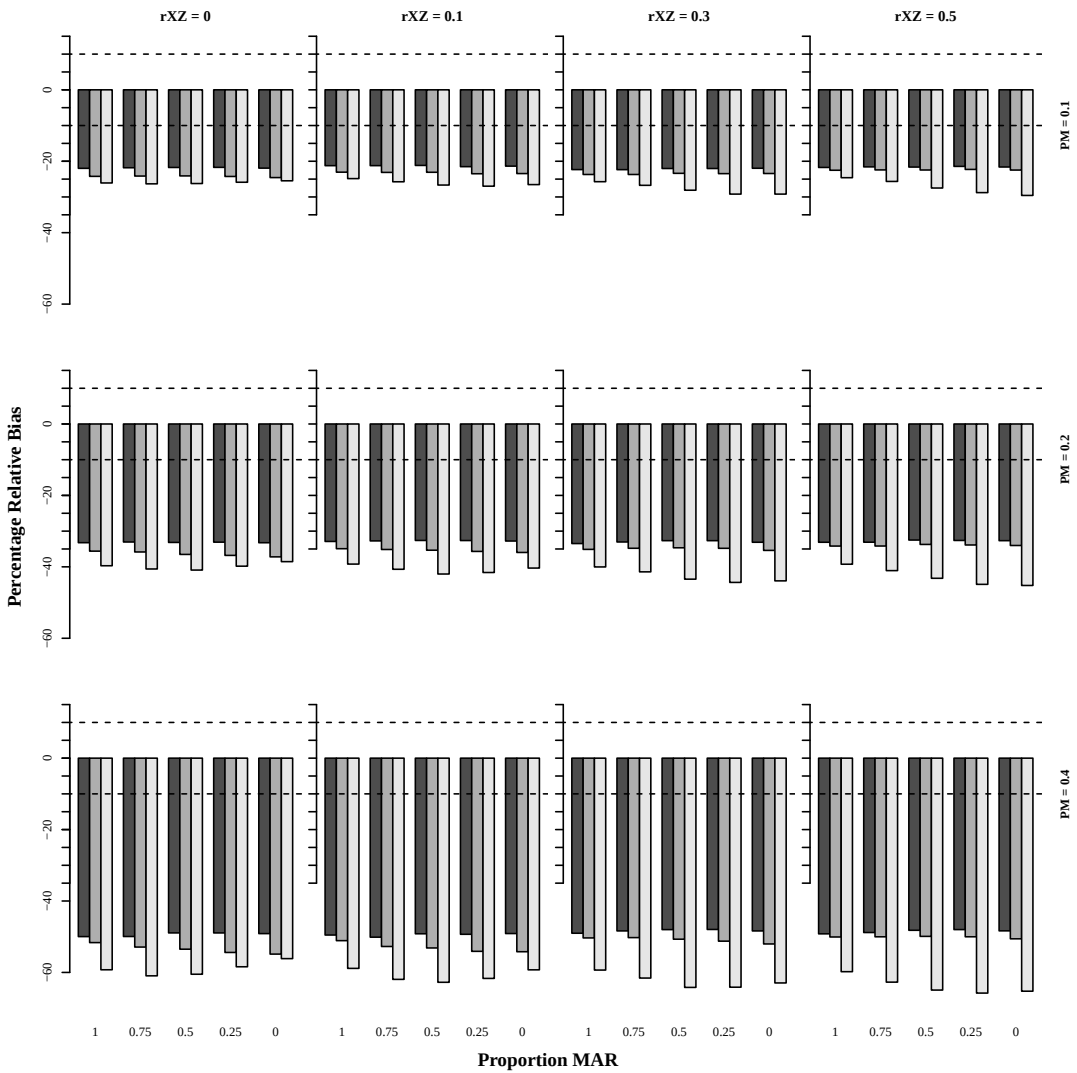
Study 1 Power with $N = 500$, $R^2 = 0.60$



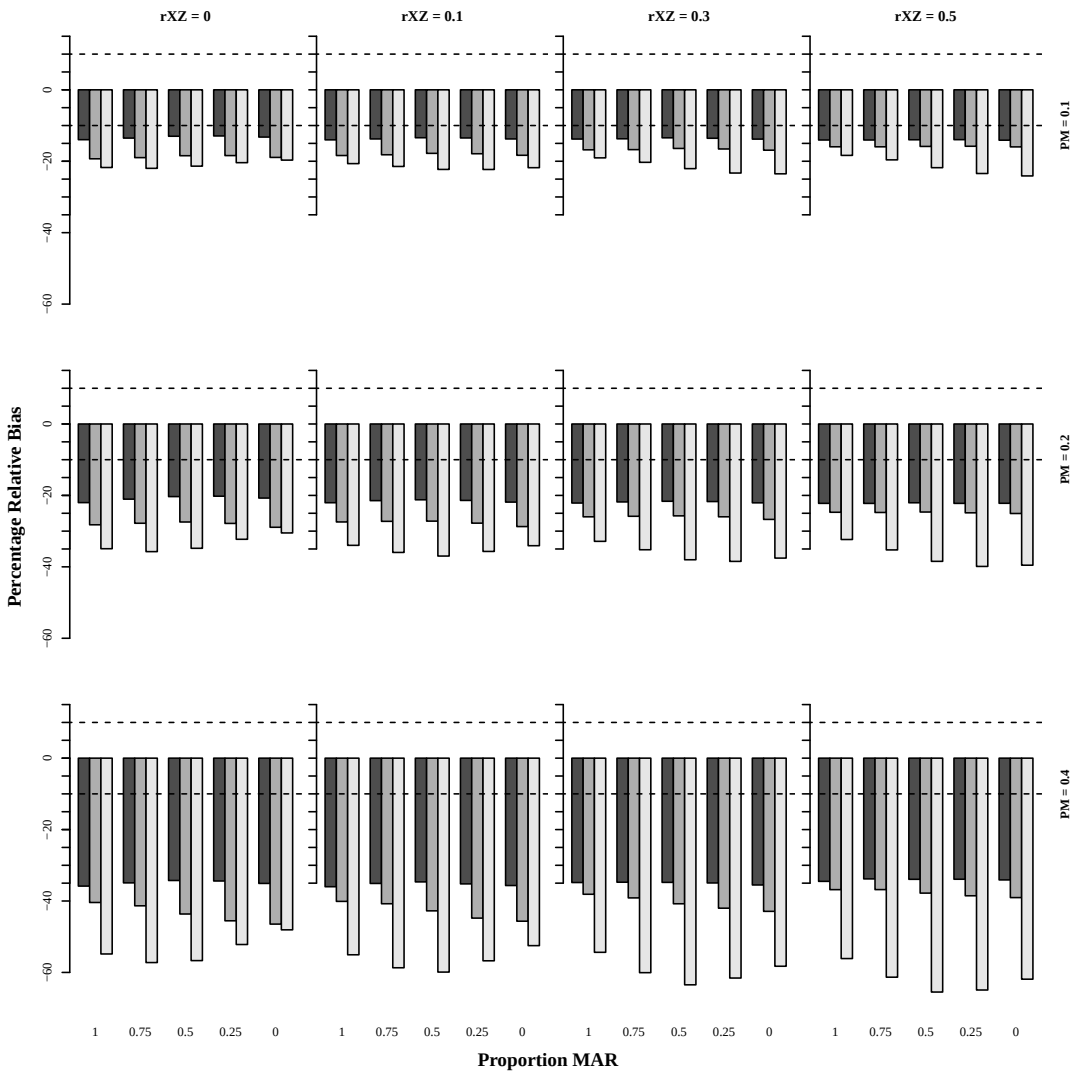
Study 2 PRB with $N = 500$, $R^2 = 0.15$



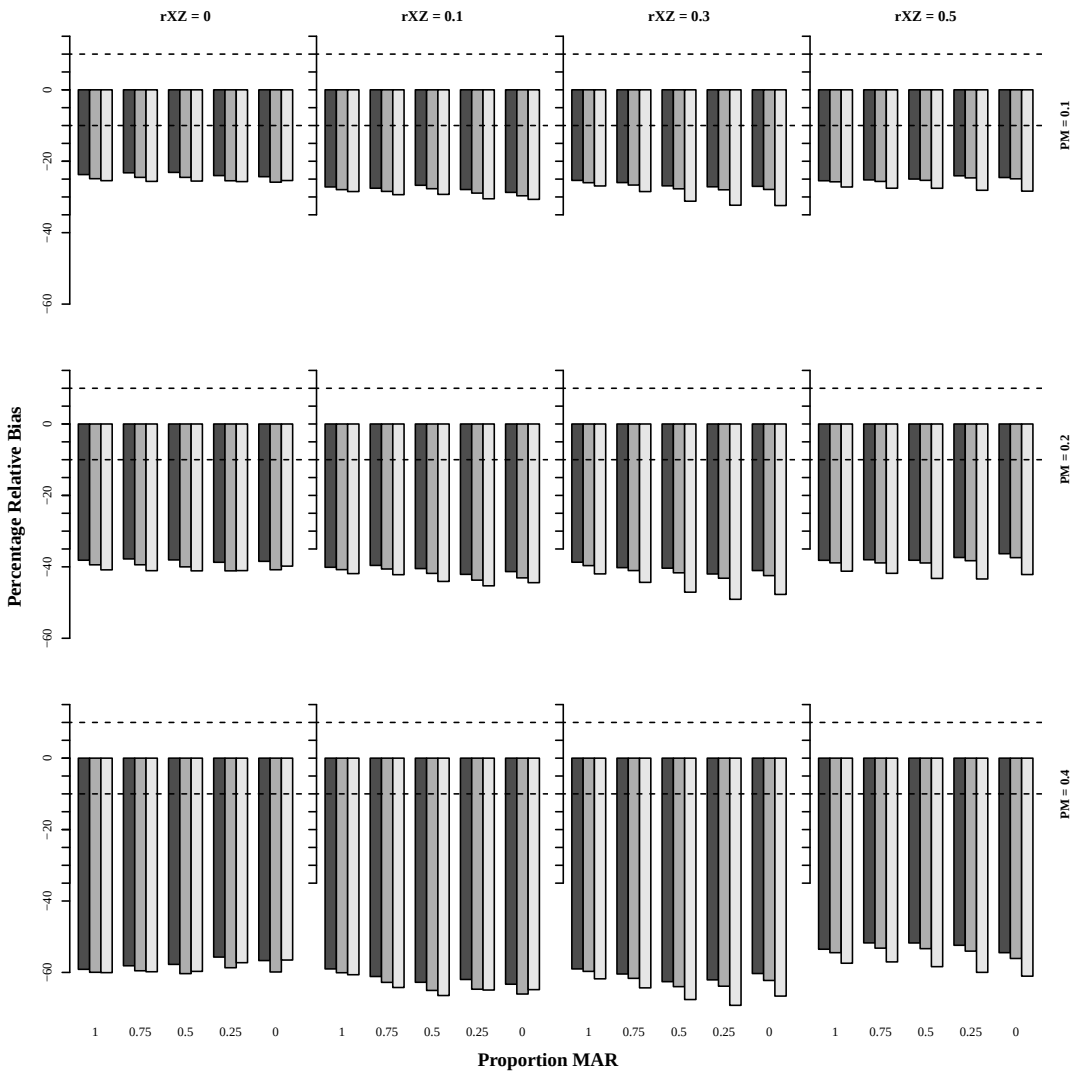
Study 2 PRB with $N = 500$, $R^2 = 0.30$



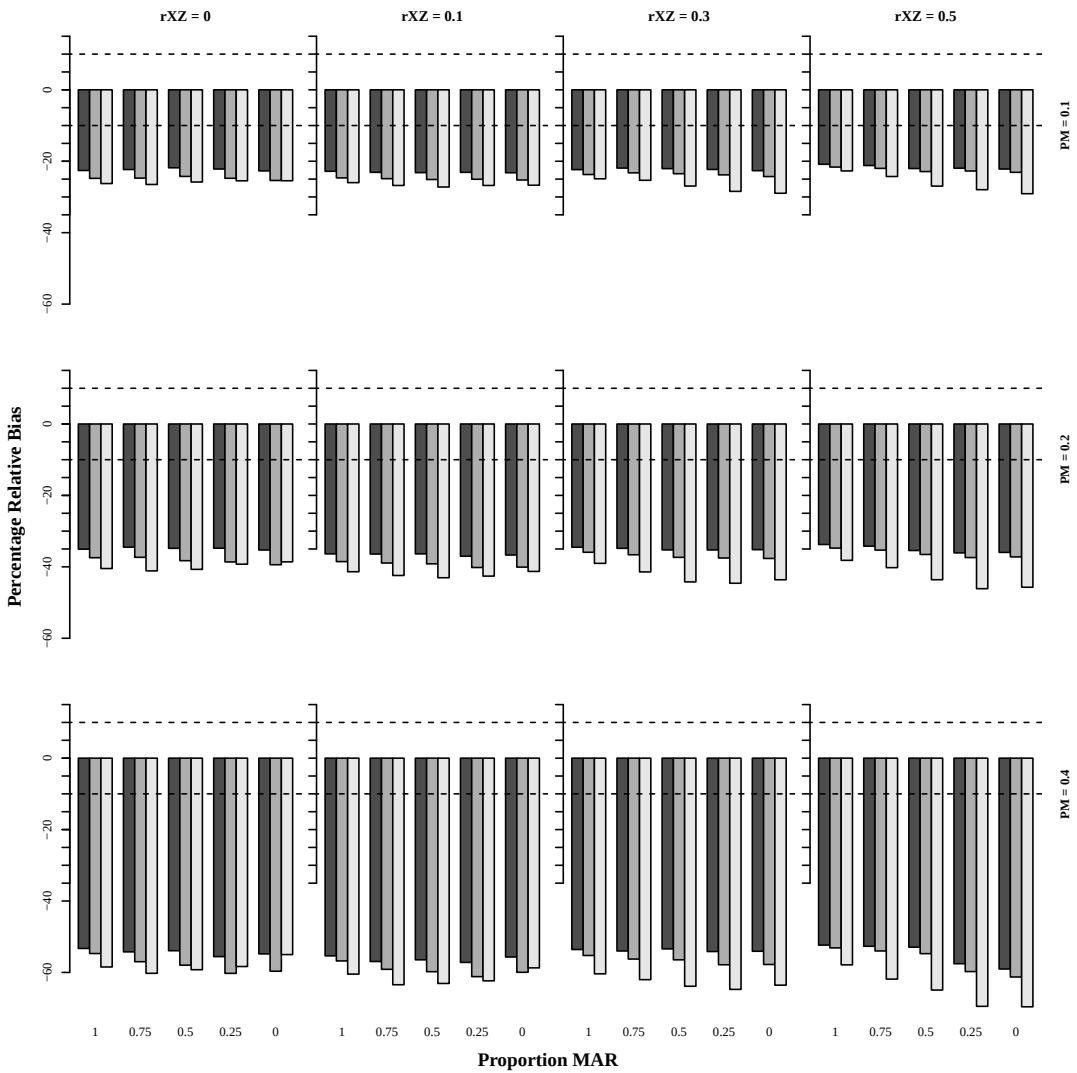
Study 2 PRB with $N = 500$, $R^2 = 0.60$



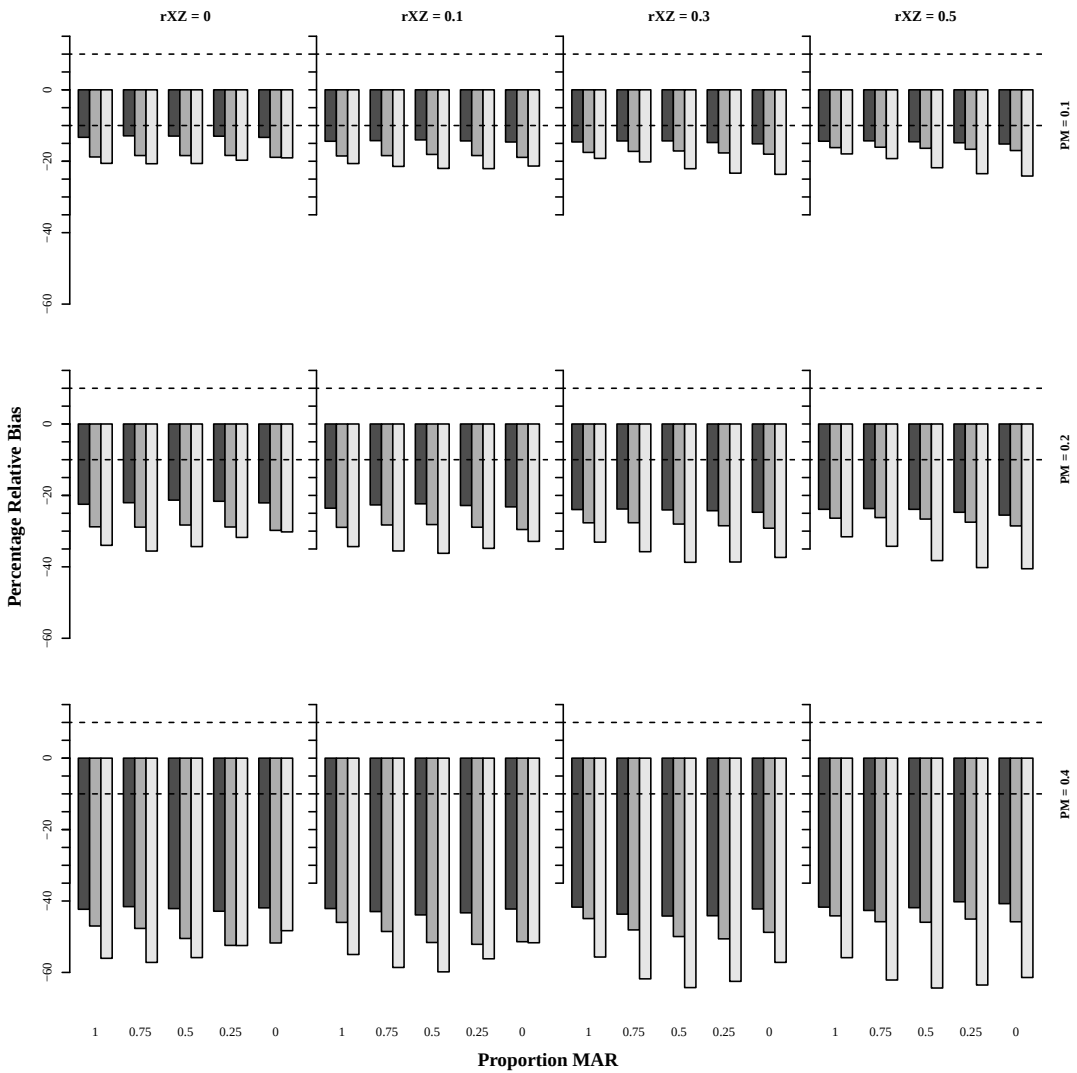
Study 2 PRB with $N = 100$, $R^2 = 0.15$



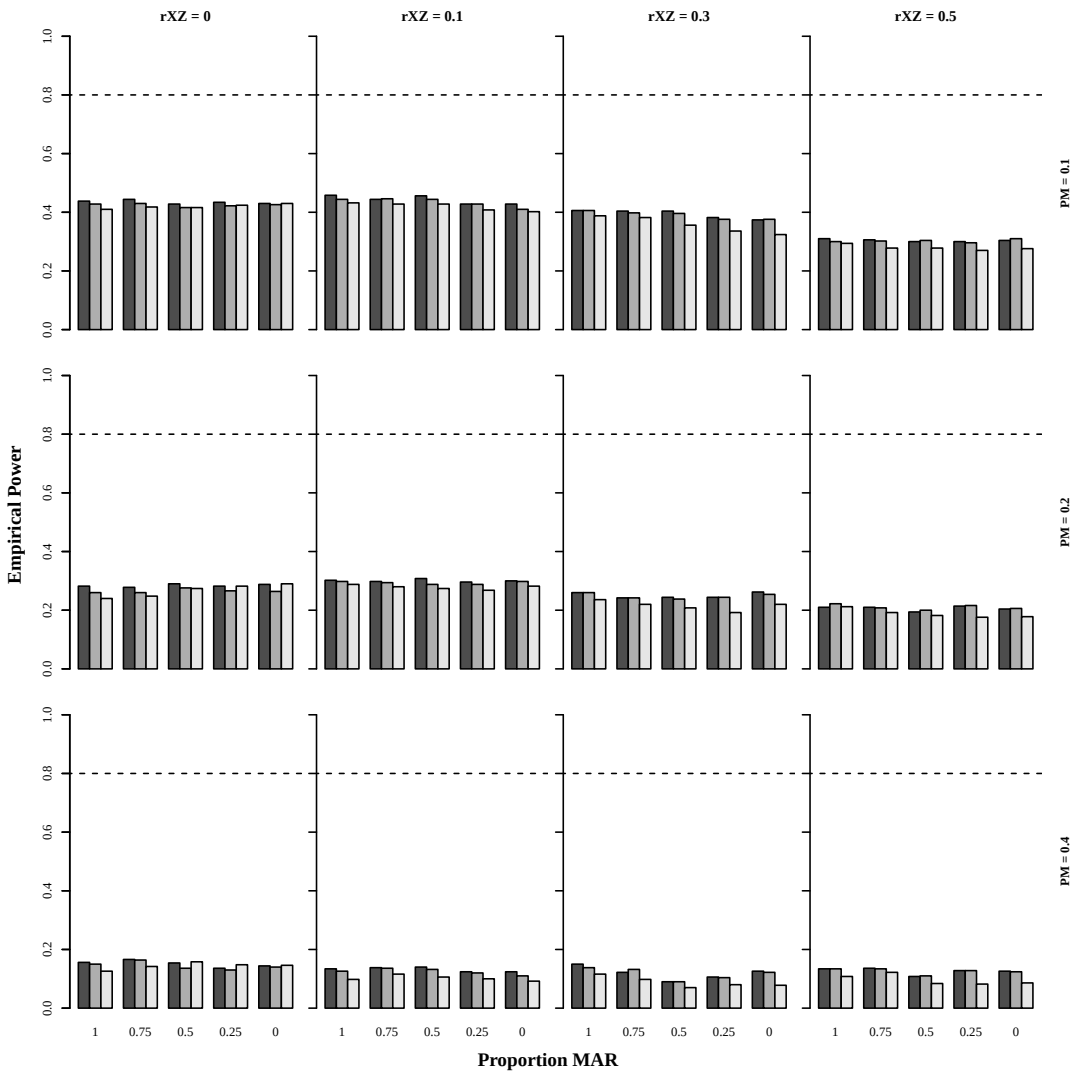
Study 2 PRB with $N = 100$, $R^2 = 0.30$



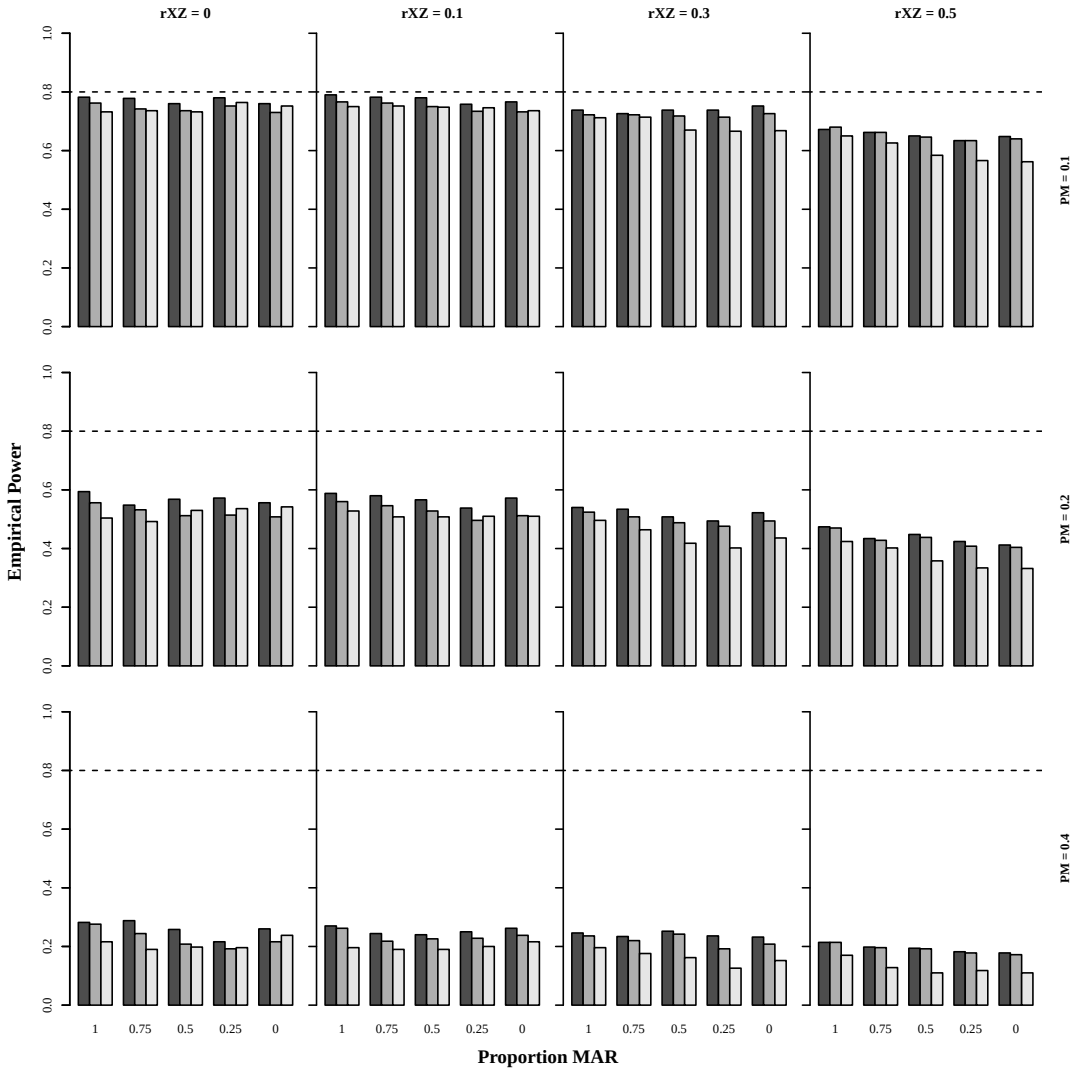
Study 2 PRB with $N = 100$, $R^2 = 0.60$



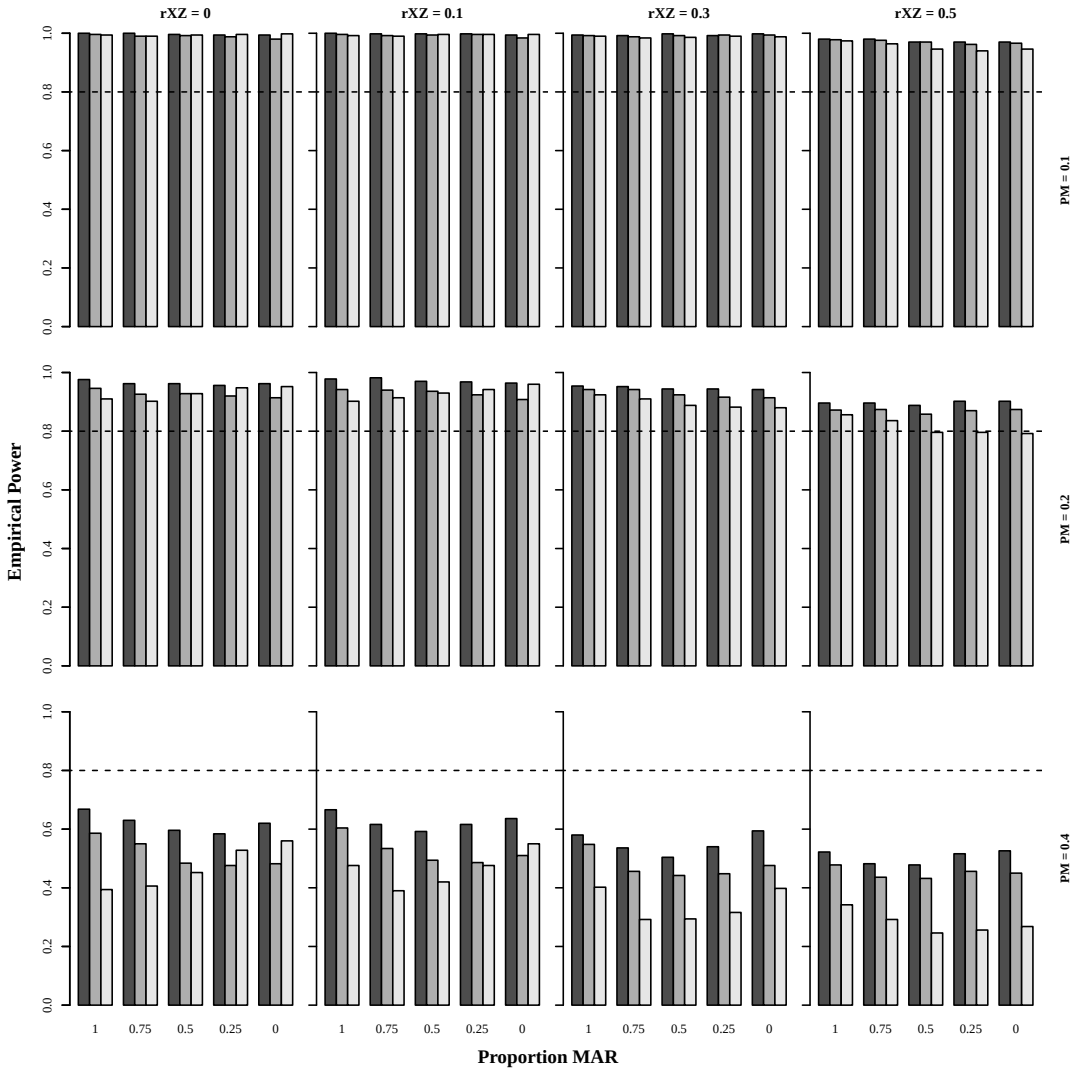
Study 2 Power with $N = 100$, $R^2 = 0.15$



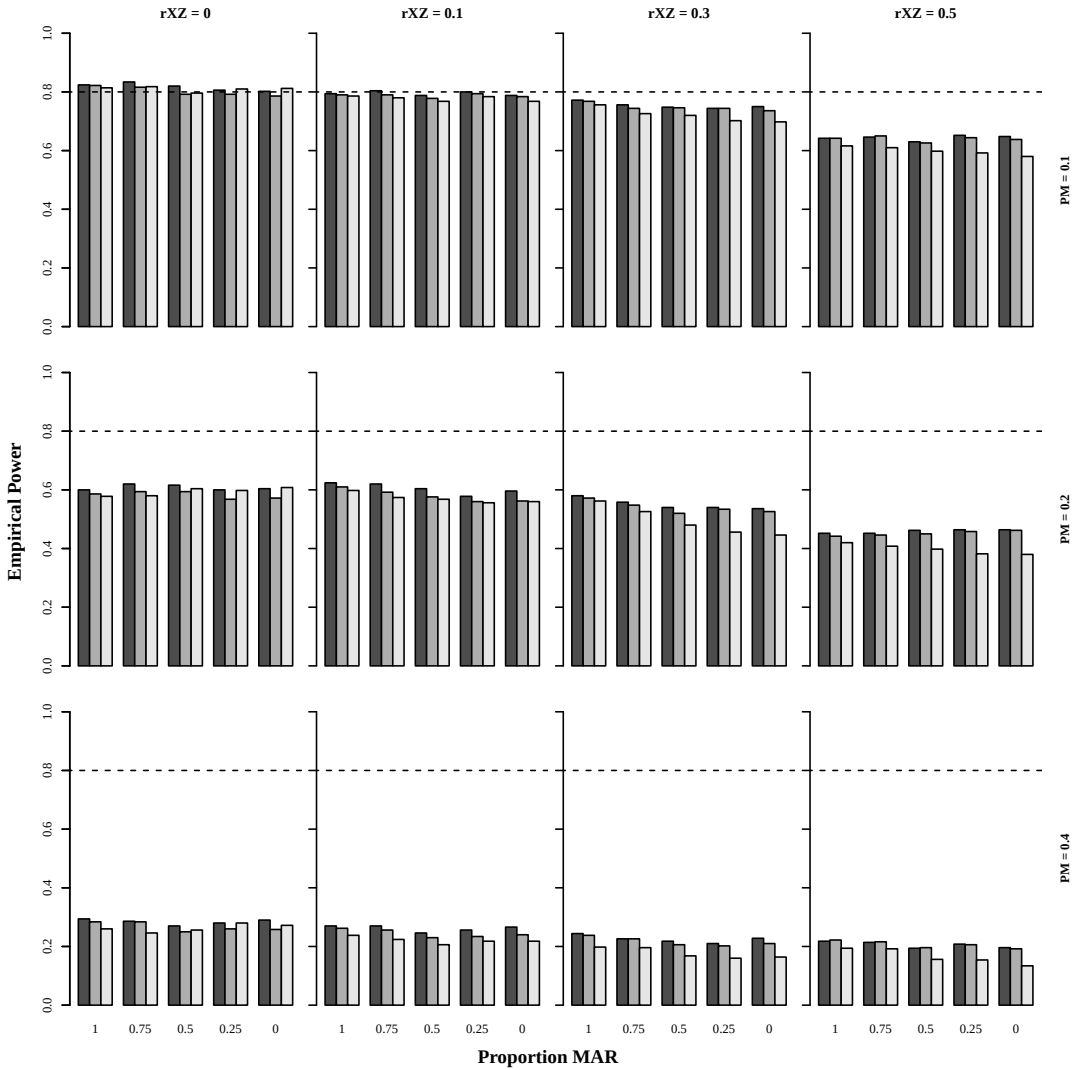
Study 2 Power with $N = 100$, $R^2 = 0.3$



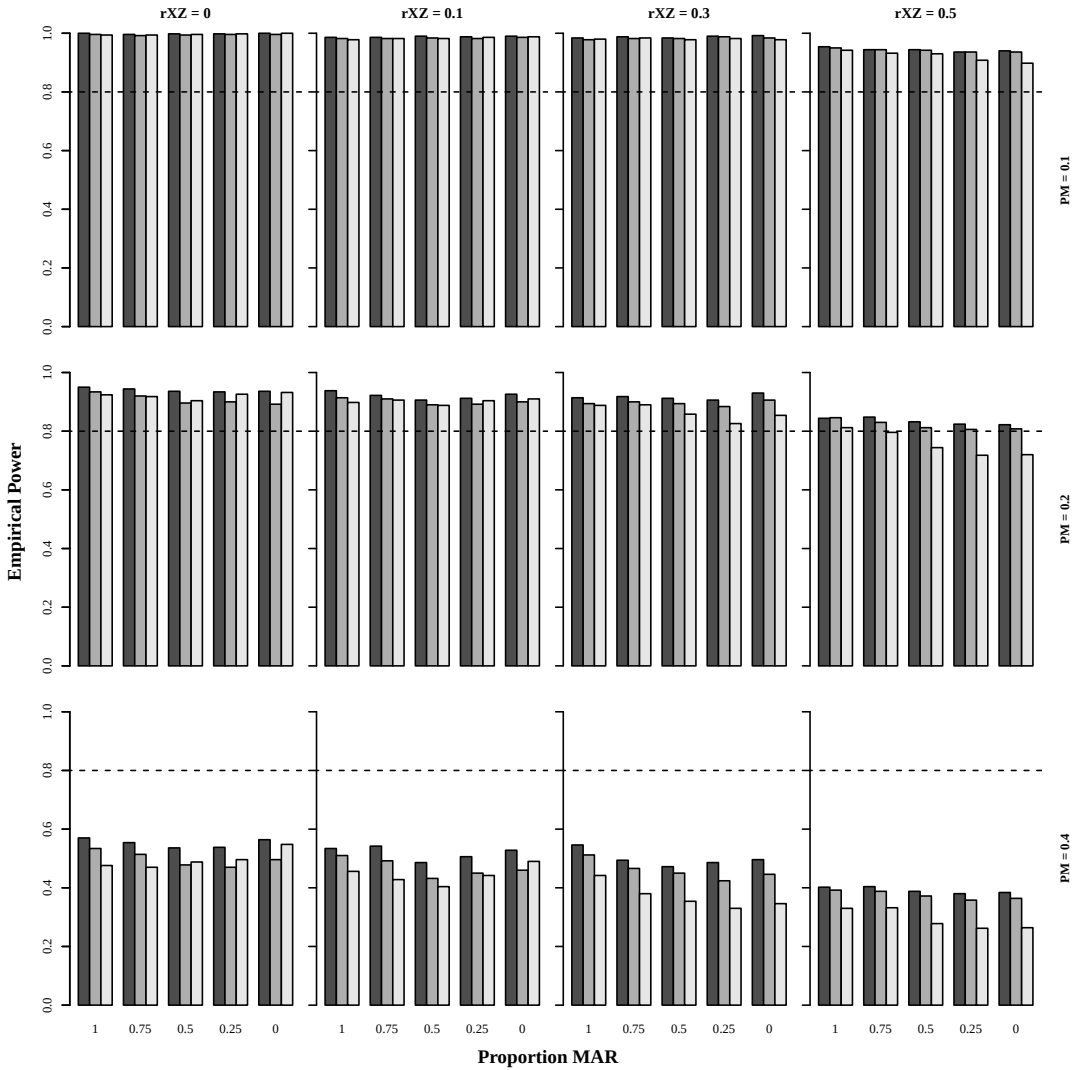
Study 2 Power with $N = 100$, $R^2 = 0.6$



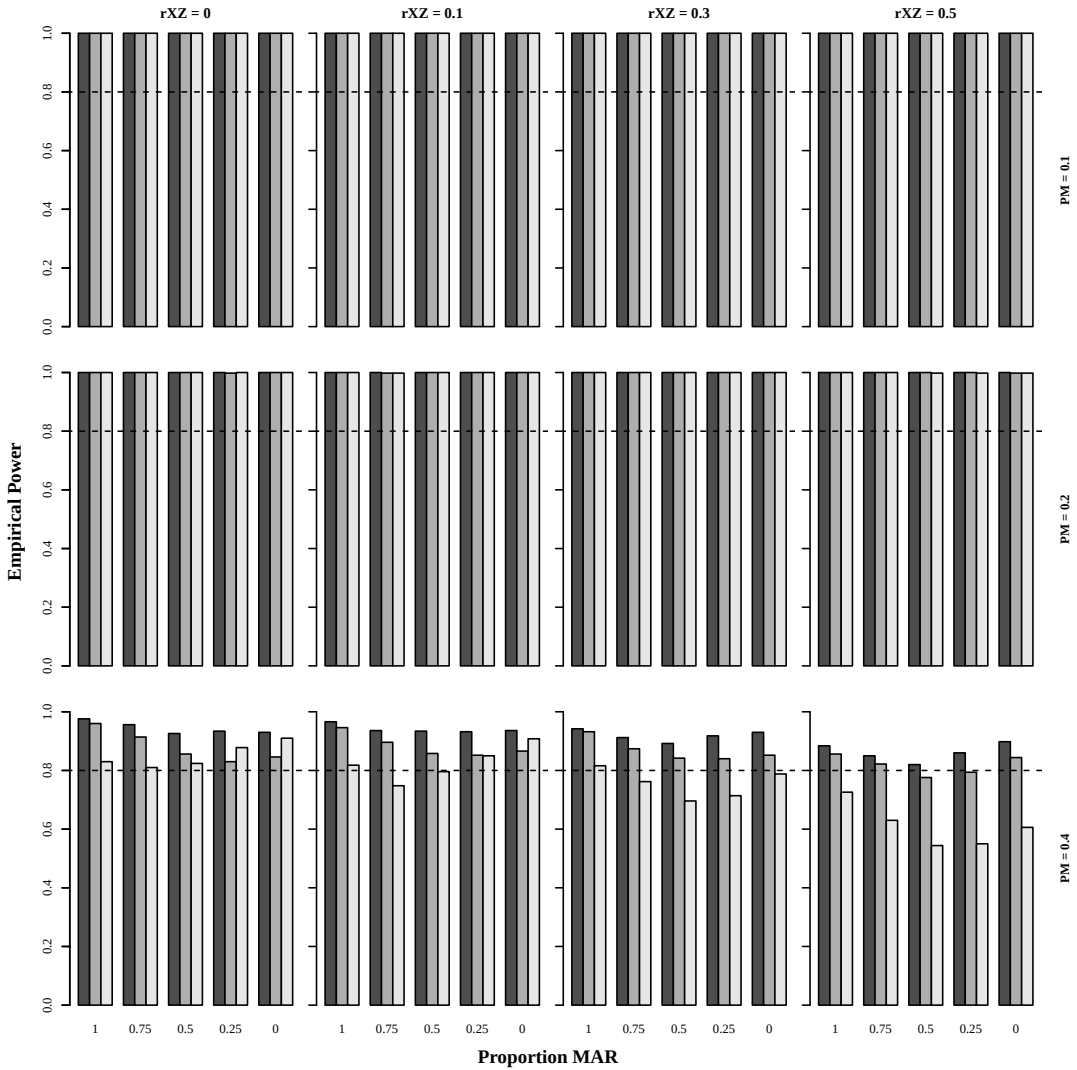
Study 2 Power with $N = 250$, $R^2 = 0.15$



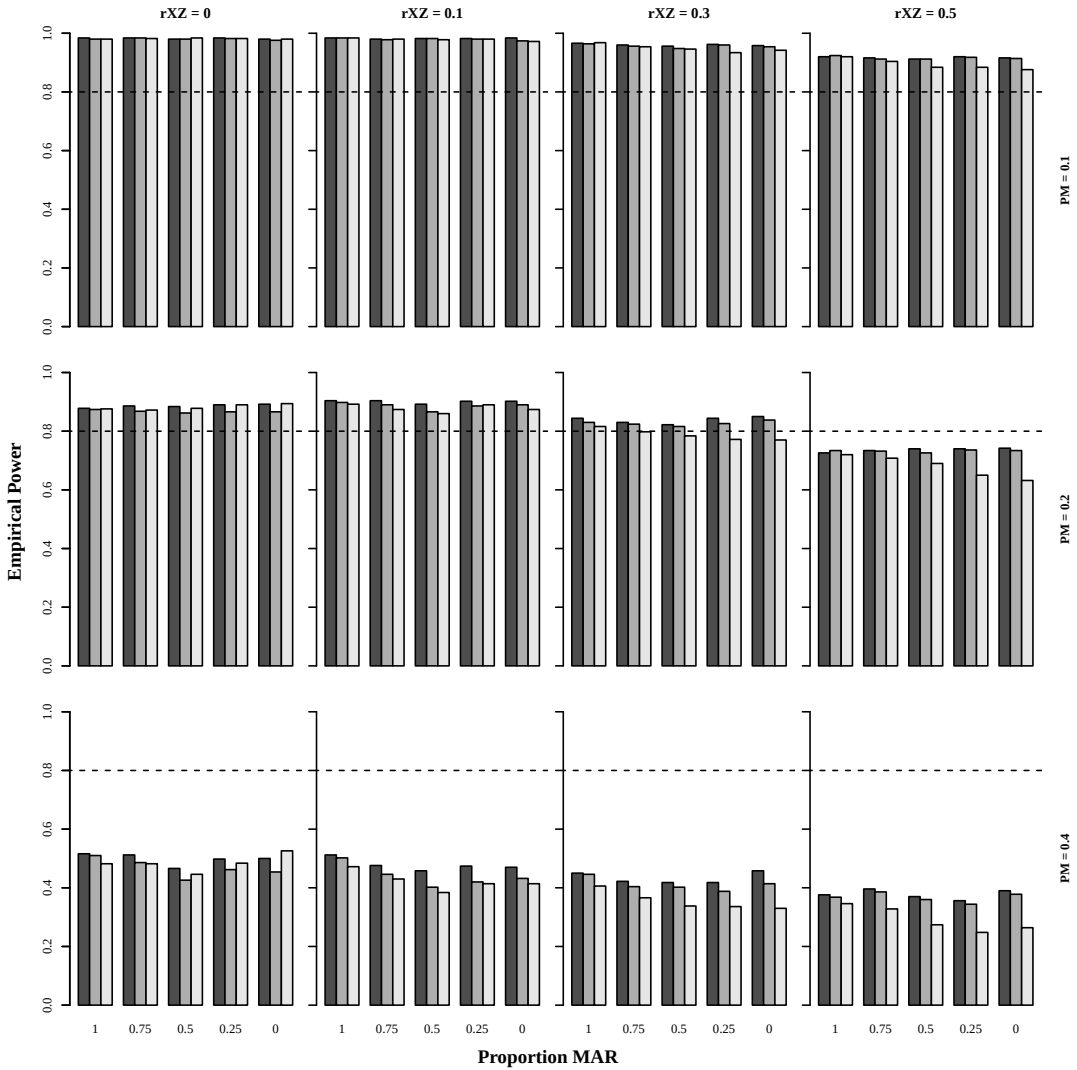
Study 2 Power with $N = 250$, $R^2 = 0.30$



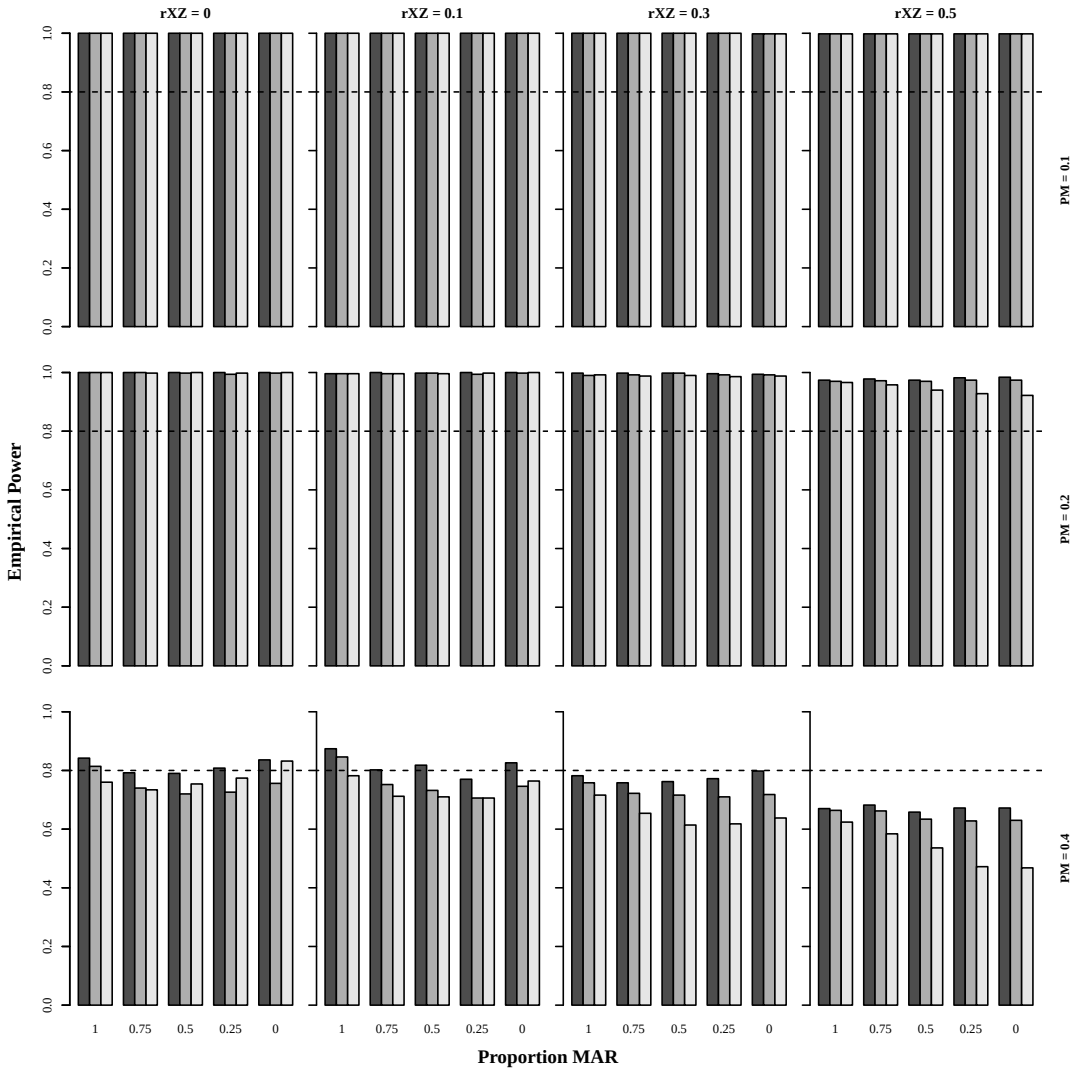
Study 2 Power with $N = 250$, $R^2 = 0.60$



Study 2 Power with $N = 500$, $R^2 = 0.15$



Study 2 Power with $N = 500$, $R^2 = 0.30$



Study 2 Power with $N = 500$, $R^2 = 0.60$