Power-Shrinkage: An Alternative Method for Dealing with Excessive Weights

Xiao-Li Meng

Joint Work with

Chihnan Chen of Center for Multicultural Mental Health Research (CMMER)

Nanhua Duan of Columbia University

and **Margarita Alegria** of Harvard University and CMMHR

“Survey weighting is a mess" (Gelman 2007, *Statistical Science*)

√ A part of the mess is that weights are all over the place!

$$\frac{Max\, Weight}{Min\, Weight} \sim 10^k, \quad k = 2, 3 \text{ or even higher.}$$

“Survey weighting is a mess" (Gelman 2007, *Statistical Science*)

- A part of the mess is that weights are all over the place!

$$\frac{Max\,Weight}{Min\,Weight} \sim 10^k, \quad k = 2, 3 \text{ or even higher.}$$

- The standard approach: trimming a small percentage of extreme weights, hoping for a smaller Mean Squared Error (MSE).

“Survey weighting is a mess" (Gelman 2007, *Statistical Science*)

- √ A part of the mess is that weights are all over the place!

$$\frac{Max\,Weight}{Min\,Weight} \sim 10^k, \quad k = 2, 3 \text{ or even higher.}$$

- √ The standard approach: trimming a small percentage of extreme weights, hoping for a smaller Mean Squared Error (MSE).
- √ The fact that the ad hoc Trimming (a.k.a "Winsorlizing") method is still a standard practice demonstrates the difficulties in dealing with weights.

Power Shrinkage

Given data and weights $\{(y_i, w_i), i = 1, \ldots, n\}$, introduce $p \in [0, 1]$, and the power-shrinkage parameter:

$$\bar{y}^{(p)} = \frac{\sum_{i=1}^{n} w_i^p y_i}{\sum_{i=1}^{n} w_i^p}.$$

Power Shrinkage

Given data and weights $\{(y_i, w_i), i = 1, \ldots, n\}$, introduce $p \in [0, 1]$, and the power-shrinkage parameter:

$$\bar{y}^{(p)} = \frac{\sum_{i=1}^{n} w_i^p y_i}{\sum_{i=1}^{n} w_i^p}.$$

- √ Clearly, $\bar{y}^{(1)}$ is the standard weighted estimator, and $\bar{y}^{(0)}$ is the unweighed estimator.
- √ The goal is to choose $p \in [0, 1]$ to achieve as smaller MSE as possible.

So Why Power Shrinkage?

- √ Survey weights are typically formed multiplicatively: $w_i = \prod_j W_{i,j}$.

So Why Power Shrinkage?

- √ Survey weights are typically formed multiplicatively: $w_i = \prod_j W_{i,j}$.
- √ Consequently, imperfections in forming the weights (e.g., measurement error) impact all weights, not just the extreme ones.

So Why Power Shrinkage?

- √ Survey weights are typically formed multiplicatively: $w_i = \prod_j W_{i,j}$.
- √ Consequently, imperfections in forming the weights (e.g., measurement error) impact all weights, not just the extreme ones.
- √ By CLT, we can often assume normality:

$$\log w_i = \sum_j \log(W_{i,j}) \sim N(A, \tau^2).$$

So Why Power Shrinkage?

- Survey weights are typically formed multiplicatively: $w_i = \prod_j W_{i,j}$.
- Consequently, imperfections in forming the weights (e.g., measurement error) impact all weights, not just the extreme ones.
- By CLT, we can often assume normality:

$$\log w_i = \sum_j \log(W_{i,j}) \sim N(A, \tau^2).$$

- Shrinking τ to $p\tau \Longleftrightarrow$ changing w to w^p.

So Why Power Shrinkage?

- Survey weights are typically formed multiplicatively: $w_i = \prod_j W_{i,j}$.
- Consequently, imperfections in forming the weights (e.g., measurement error) impact all weights, not just the extreme ones.
- By CLT, we can often assume normality:

$$\log w_i = \sum_j \log(W_{i,j}) \sim N(A, \tau^2).$$

- Shrinking τ to $p\tau \iff$ changing w to w^p.
- Using w_i^p preserves the (strict) order of w_i's.

The Most Frequently Asked Question By Referees

- √ ***"So how do you choose p?"***

The Most Frequently Asked Question By Referees

- ***"So how do you choose p?"***
- We seek p to reduce MSE, compared to both $\bar{y}^{(1)}$ and $\bar{y}^{(0)}$.

The Most Frequently Asked Question By Referees

- ***"So how do you choose p?"***
- We seek p to reduce MSE, compared to both $\bar{y}^{(1)}$ and $\bar{y}^{(0)}$.
- Intuitively, $p = 1/2$ seems to be a good compromise: what else could it be?

The Most Frequently Asked Question By Referees

- ***"So how do you choose p?"***
- We seek p to reduce MSE, compared to both $\bar{y}^{(1)}$ and $\bar{y}^{(0)}$.
- Intuitively, $p = 1/2$ seems to be a good compromise: what else could it be?
- $p = 1/2$ was suggested in a thesis by Levenson (1993, U of Chicago) in the context of Importance Sampling.

The Most Frequently Asked Question By Referees

- ***"So how do you choose p?"***
- We seek p to reduce MSE, compared to both $\bar{y}^{(1)}$ and $\bar{y}^{(0)}$.
- Intuitively, $p = 1/2$ seems to be a good compromise: what else could it be?
- $p = 1/2$ was suggested in a thesis by Levenson (1993, U of Chicago) in the context of Importance Sampling.
- Empirically, I was reminded by my twin brother ...

Evidence of Taking Root from Gelman and Little (1998)?

Table 2. Poststratification Weights for Late CBS Polls, Early CBS Polls, and NES, Normalized So That the Weight is 1 for Respondents from Households with One Adult

Number of Adults in Household	Poststratification Weights			
	Theory	Early CBS	Late CBS	NES
1	1	1.00	1.00	1.00
2	2	1.32	1.38	2.00
3	3	1.35	1.53	2.30
4+	4.25	0.95	1.20	2.55

NOTE.—If sampling all went as planned, the weights would equal the theoretical values. (The last weight is not exactly 4 because the last poststratification category includes all households with 4 or more adults.) The weights for the higher categories are lower than the theoretical values because the surveys oversampled the larger households.

Potential Thesis Topics

- √ Is it possible to prove that $p = 1/2$ is some kind of minimax choice?

Potential Thesis Topics

- √ Is it possible to prove that $p = 1/2$ is some kind of minimax choice?
- √ Is it possible to find theoretically the optimal p?

Potential Thesis Topics

- ✓ Is it possible to prove that $p = 1/2$ is some kind of minimax choice?
- ✓ Is it possible to find theoretically the optimal p?
- ✓ Is it possible to estimate such an optimal p?

Potential Thesis Topics

- √ Is it possible to prove that $p = 1/2$ is some kind of minimax choice?
- √ Is it possible to find theoretically the optimal p?
- √ Is it possible to estimate such an optimal p?
- √ We cannot solve any of these problems, because all of us already have Ph.D degrees.

Potential Thesis Topics

- √ Is it possible to prove that $p = 1/2$ is some kind of minimax choice?
- √ Is it possible to find theoretically the optimal p?
- √ Is it possible to estimate such an optimal p?
- √ We cannot solve any of these problems, because all of us already have Ph.D degrees.
- √ So we took an easier route ...

Conducting Simulation Studies via CPES

- √ The Collaborative Psychiatric Epidemiology Surveys (CPES) were initiated in recognition of the factors of mental disorders among the general population with emphasis on minority groups.

Conducting Simulation Studies via CPES

- The Collaborative Psychiatric Epidemiology Surveys (CPES) were initiated in recognition of the factors of mental disorders among the general population with emphasis on minority groups.
- It combines three (hopefully!) nationally representative multi-stage surveys: the National Comorbidity Survey Replication (NCS-R), the National Survey of American Life (NSAL), and the National Latino and Asian American Study (NLAAS).

Conducting Simulation Studies via CPES

- √ The Collaborative Psychiatric Epidemiology Surveys (CPES) were initiated in recognition of the factors of mental disorders among the general population with emphasis on minority groups.
- √ It combines three (hopefully!) nationally representative multi-stage surveys: the National Comorbidity Survey Replication (NCS-R), the National Survey of American Life (NSAL), and the National Latino and Asian American Study (NLAAS).
- √ Data were collected between May 2002 and November 2003, resulting 13,837 cases.

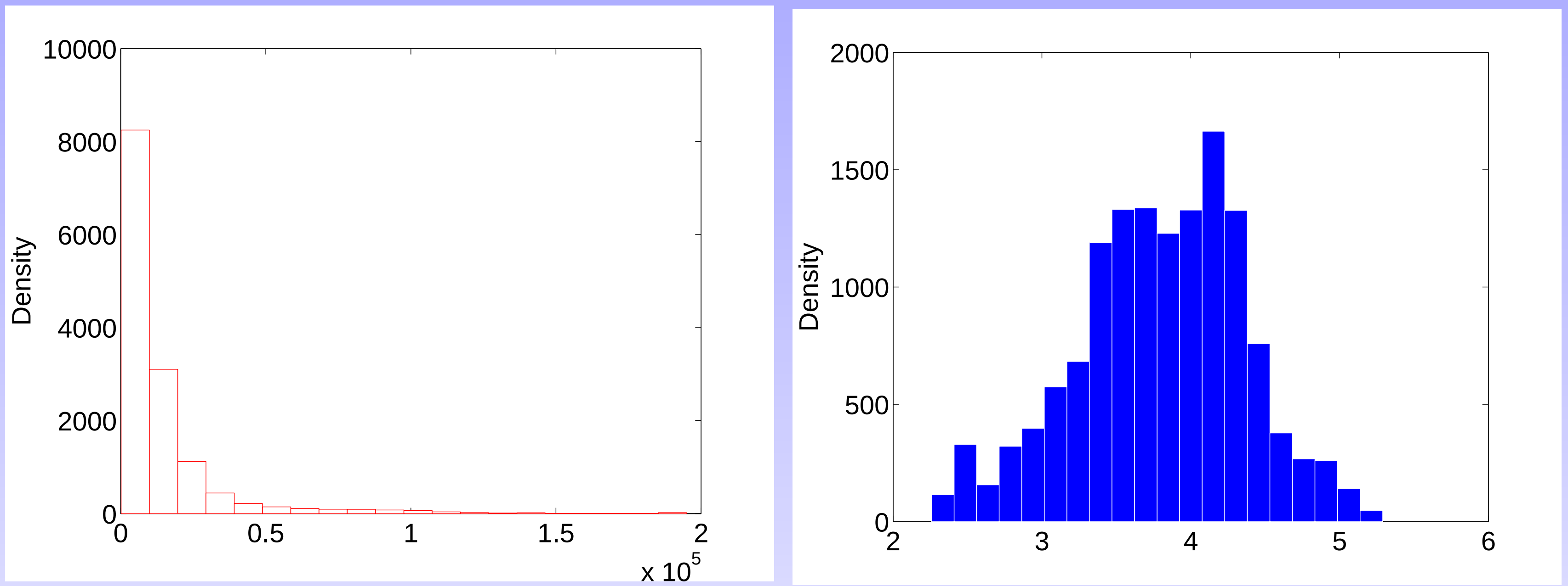
How Variable are the Survey Weights?

	w	Percentile	w
MIN	181.5	5	550.6
MEAN	13,496.1	25	2,722.7
MAX	195,000.0	50	6,799.1
		75	15,942.0
		95	50,167.1

Table 1: Summery Statistics for w

Always Take Log!

Survey Weight w $\qquad\qquad$ $\log_{10}(w)$



Love Log or Log for Love!

"Can we log on together?"

Create a Semi-artificial Population

- Each of the $I = 13,837$ reported y_i represents a cluster with w_i individuals, whose values (if continuous and no restriction) are generated according to $N(y_i, \sigma_y^2)$, where σ_y is the standard error of y. The population consisted of $N = \sum_{i=1}^{I} w_i = 186,745,266$ individuals.

Create a Semi-artificial Population

- Each of the $I = 13,837$ reported y_i represents a cluster with w_i individuals, whose values (if continuous and no restriction) are generated according to $N(y_i, \sigma_y^2)$, where σ_y is the standard error of y. The population consisted of $N = \sum_{i=1}^{I} w_i = 186,745,266$ individuals.
- If y has to be positive, a log-normal distribution is used instead.

Create a Semi-artificial Population

- ✓ Each of the $I = 13,837$ reported y_i represents a cluster with w_i individuals, whose values (if continuous and no restriction) are generated according to $N(y_i, \sigma_y^2)$, where σ_y is the standard error of y. The population consisted of $N = \sum_{i=1}^{I} w_i = 186,745,266$ individuals.
- ✓ If y has to be positive, a log-normal distribution is used instead.
- ✓ If y is binary, we use a logistic model with age, gender, height and education as covariates, and then sample from the Bernoulli distribution with the predicted mean. (For gender, we use age, height and education as covariates.)

Simulation Design

- √ Two-stage Sampling:

Simulation Design

- √ Two-stage Sampling:
  - ⋆ First, draw q clusters by simple random sampling without replacement.

Simulation Design

√ Two-stage Sampling:

- ⋆ First, draw q clusters by simple random sampling without replacement.
- ⋆ Second, draw s cases within each cluster by simple random sampling without replacement.

Simulation Design

- Two-stage Sampling:
  - First, draw q clusters by simple random sampling without replacement.
  - Second, draw s cases within each cluster by simple random sampling without replacement.
- The simulated sample size is $n = q \times s$. The observation from cluster i is assigned the weight $w_i^* = \frac{w_i}{s} \propto w_i$.

Simulation Design

- Two-stage Sampling:
  - First, draw q clusters by simple random sampling without replacement.
  - Second, draw s cases within each cluster by simple random sampling without replacement.
- The simulated sample size is $n = q \times s$. The observation from cluster i is assigned the weight $w_i^* = \frac{w_i}{s} \propto w_i$.
- For the reported results, $s = 2$.

Variables Examined

- √ Gender, age, height, household income, major depression, substance disorder, social phobia, any disorder, ageplusswgt, body weight, nativity and the survey weight w itself.

Variables Examined

- √ Gender, age, height, household income, major depression, substance disorder, social phobia, any disorder, agepluswgt, body weight, nativity and the survey weight w itself.
- √ ageplusw gt $= \text{age} + 0.001 * w$, which correlates with w with *unweighted* correlation $r = 0.5360$.

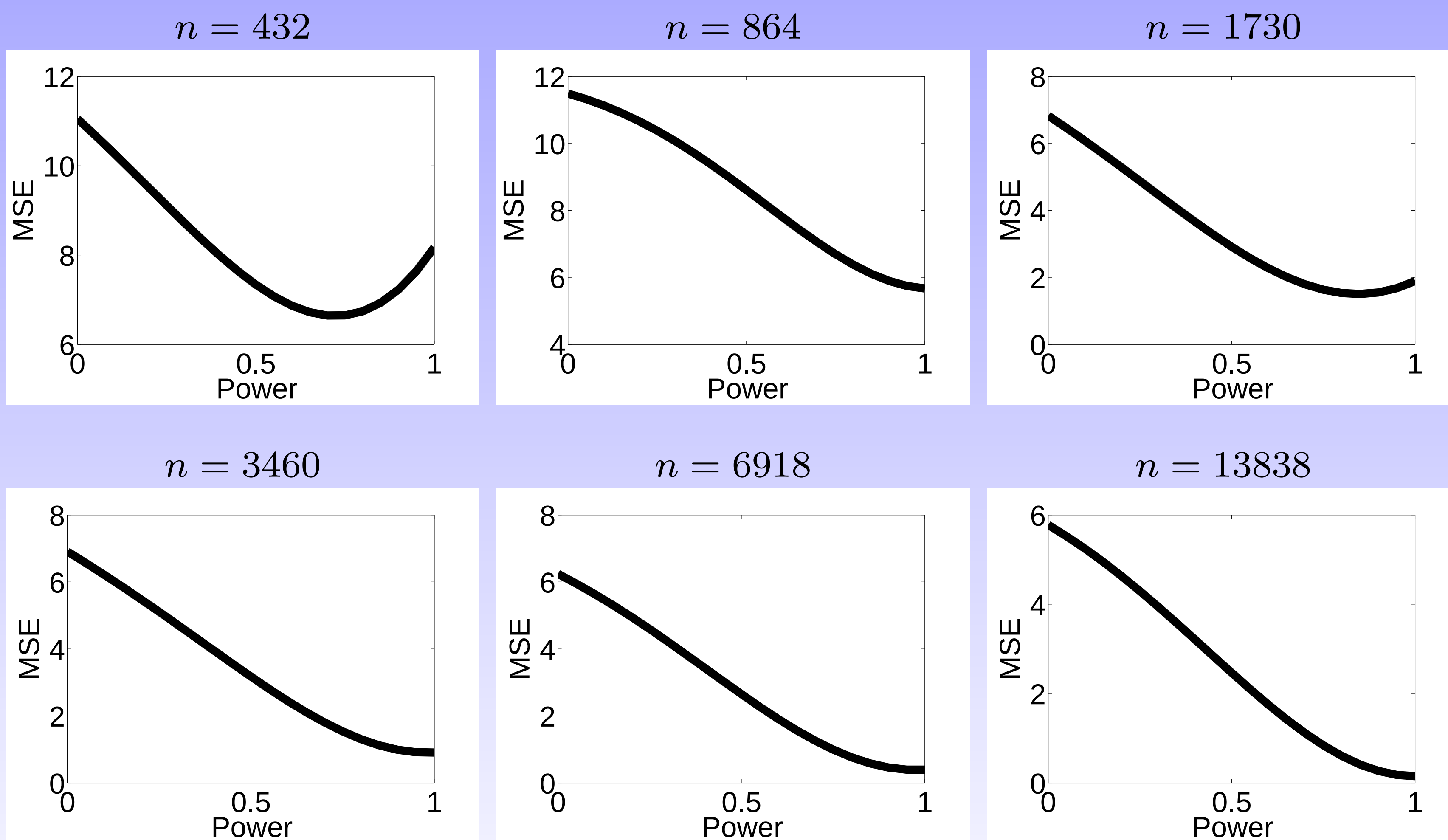
Variables Examined

- Gender, age, height, household income, major depression, substance disorder, social phobia, any disorder, ageplusswgt, body weight, nativity and the survey weight w itself.
- ageplusswgt $=$ age $+ 0.001 * w$, which correlates with w with *unweighted* correlation $r = 0.5360$.
- The importance of considering unweighted correlation:

$$\bar{y}^{(1)} - \bar{y}^{(0)} = \frac{\sum_i w_i y_i}{\sum_i w_i} - \frac{\sum_i y_i}{n} = \frac{\mathrm{Cov_n}(w, y)}{\bar{w}}$$

Age

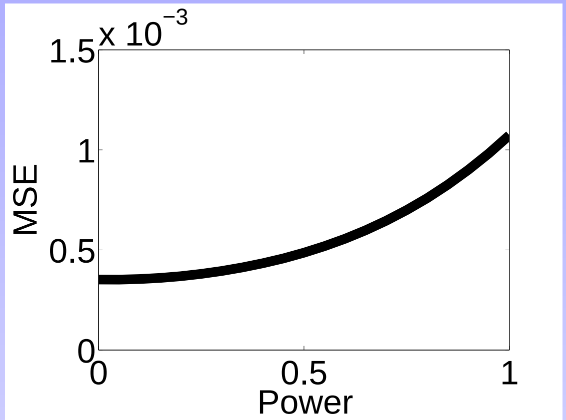
$$\bar{r} = 0.0617$$



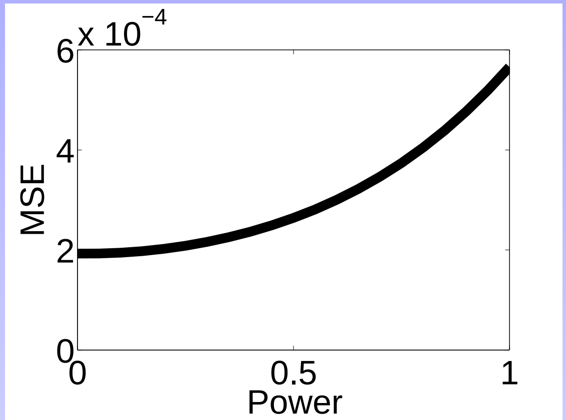
Major Depression

$$\bar{r} = \text{-0.0069}$$

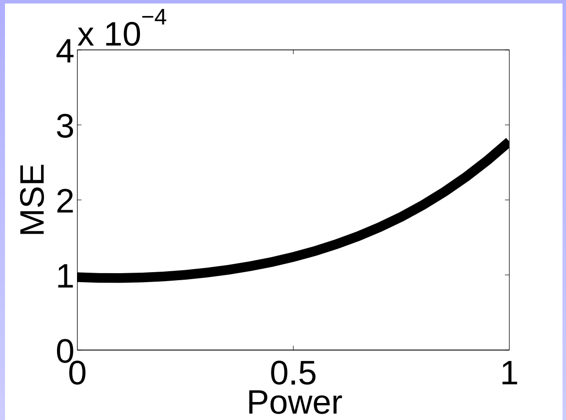
$n = 432$



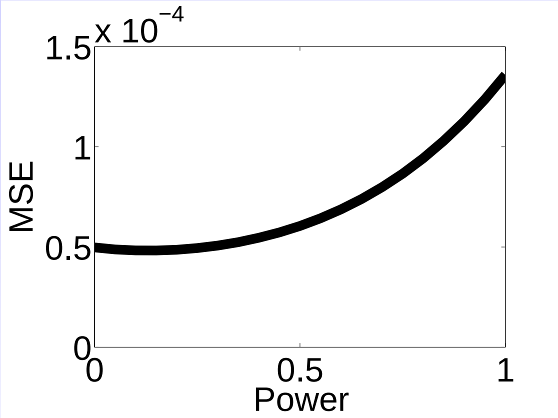
$n = 864$



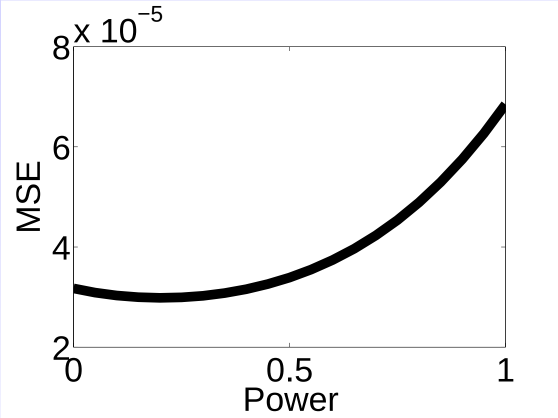
$n = 1730$



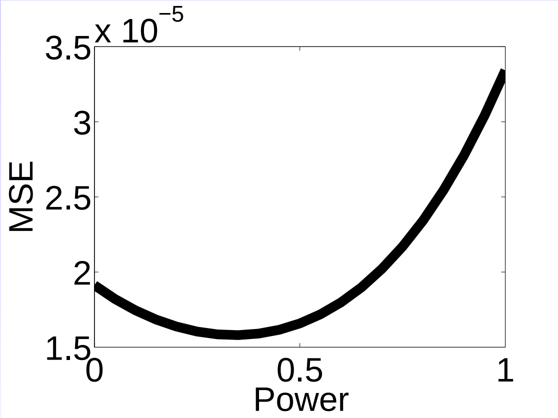
$n = 3460$



$n = 6918$

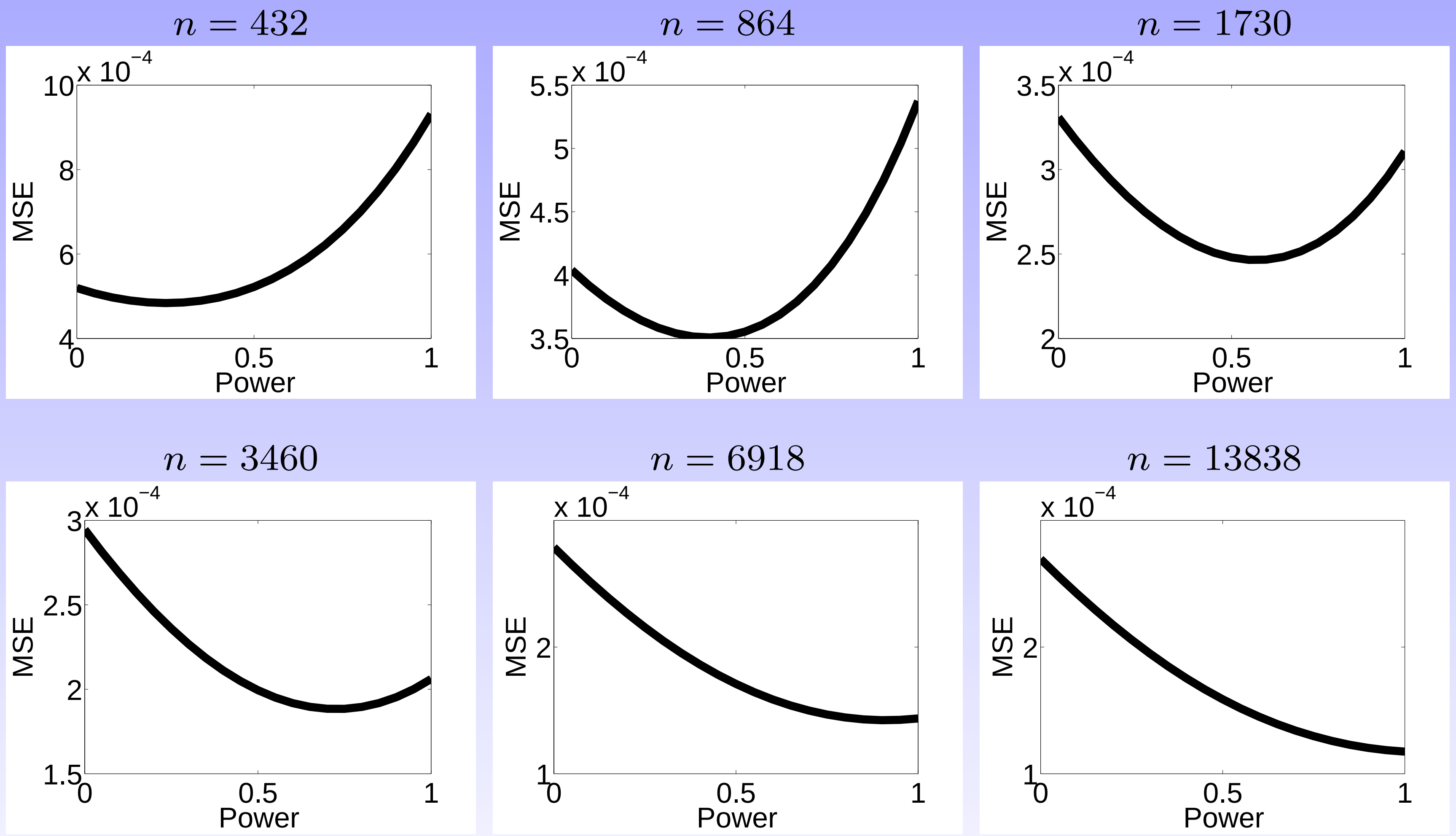


$n = 13838$



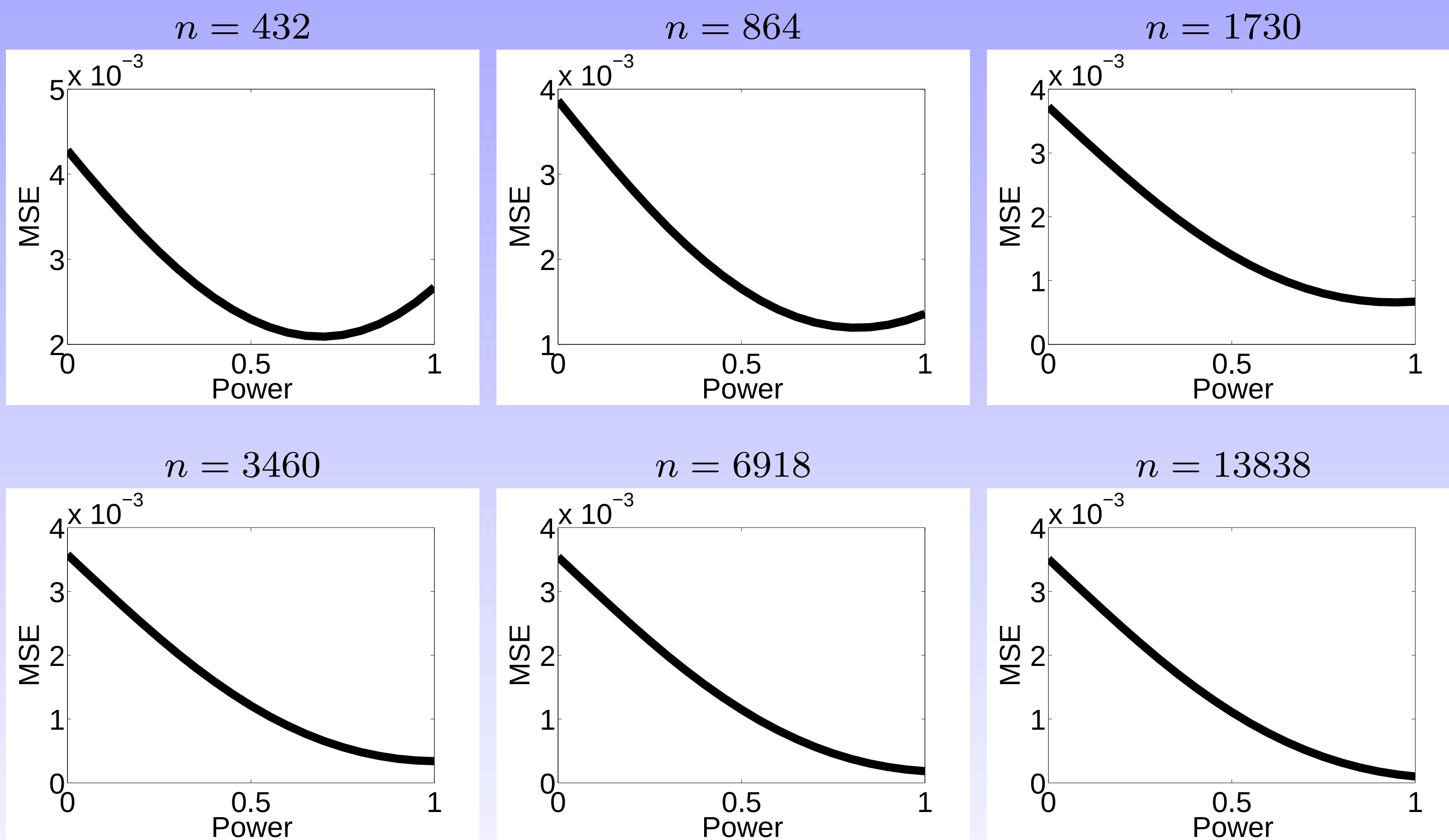
Substance Abuse

$$\bar{r} = 0.0139$$



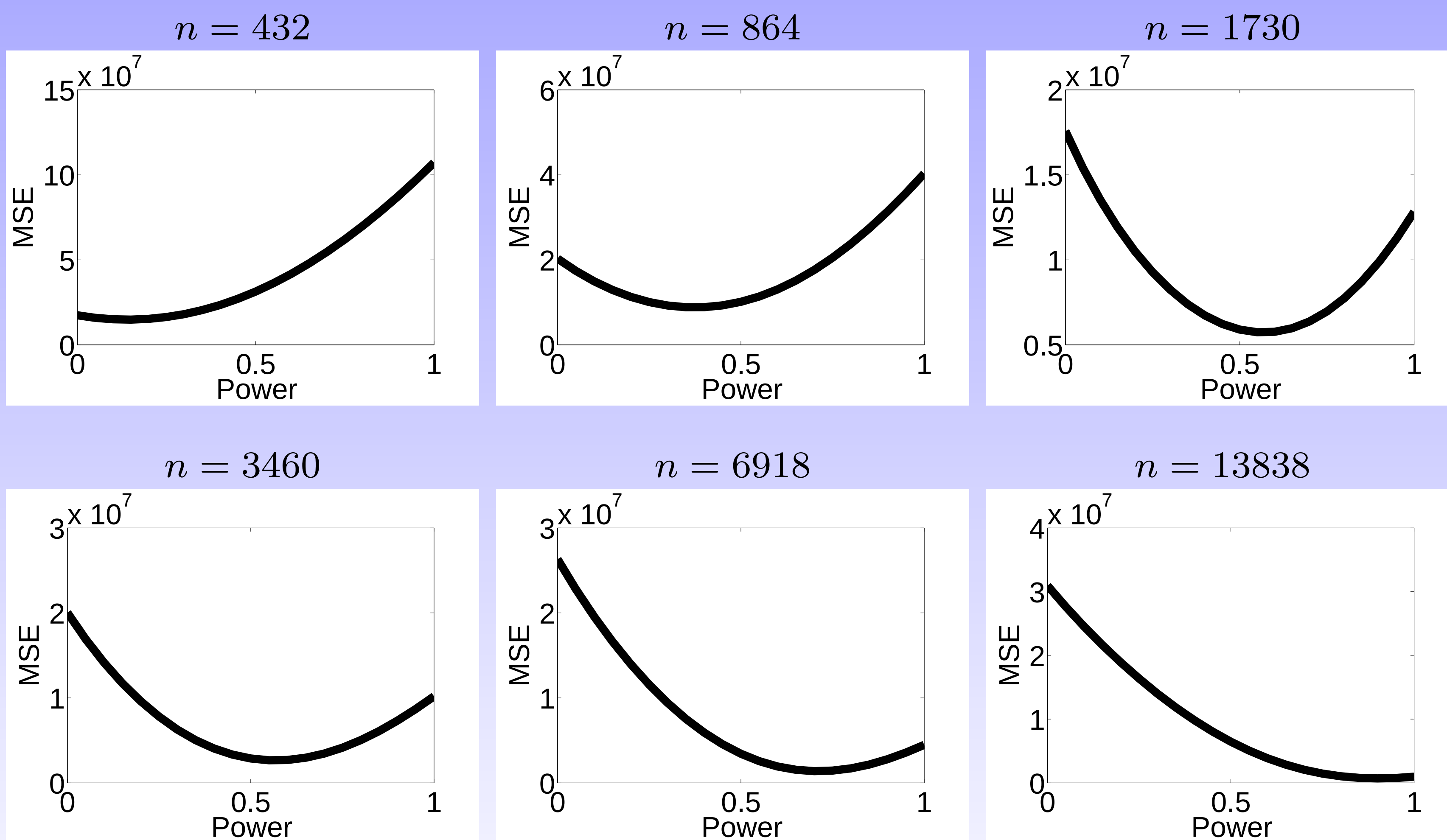
Gender

$$\bar{r} = \text{-0.0716}$$



Household Income

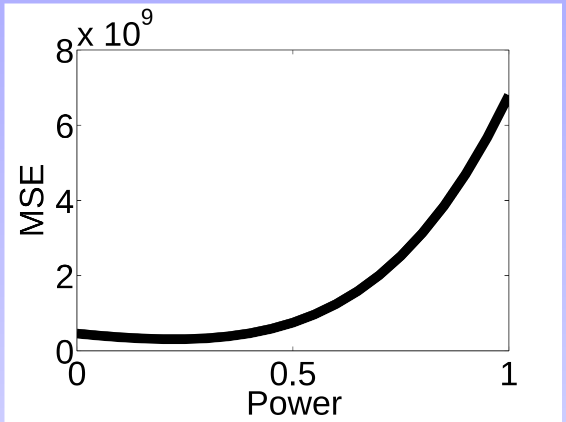
$$\bar{r} = 0.0809$$



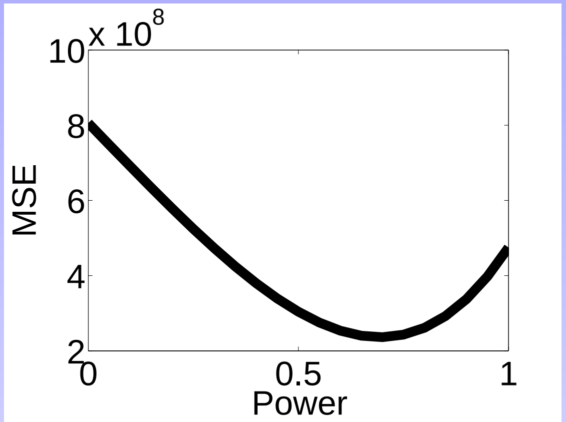
Survey Weight w

$$\bar{r} = 0.1687$$

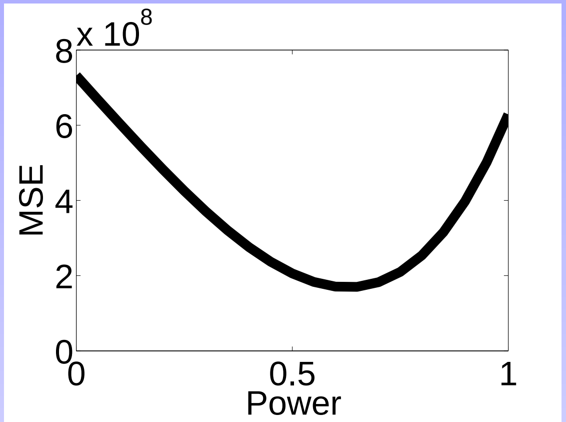
$n = 432$



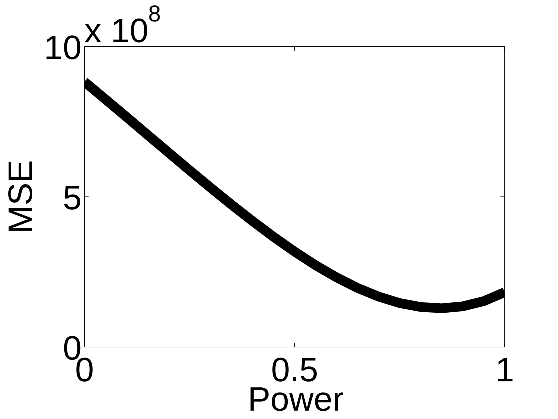
$n = 864$



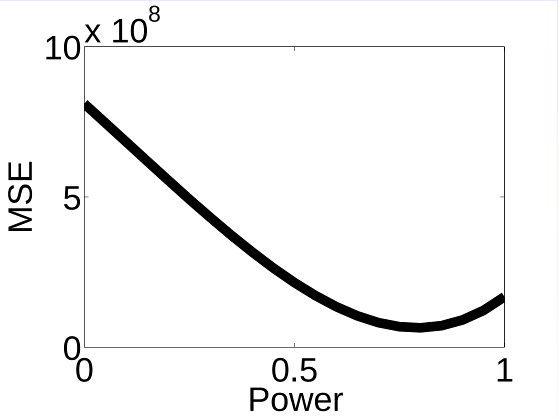
$n = 1730$



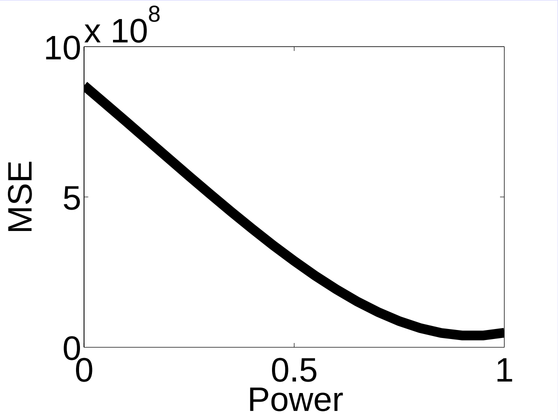
$n = 3460$



$n = 6918$



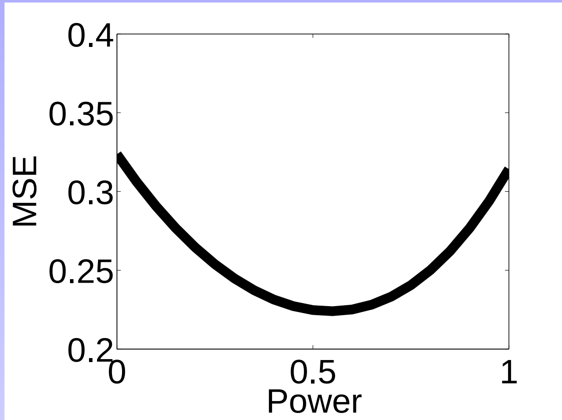
$n = 13838$



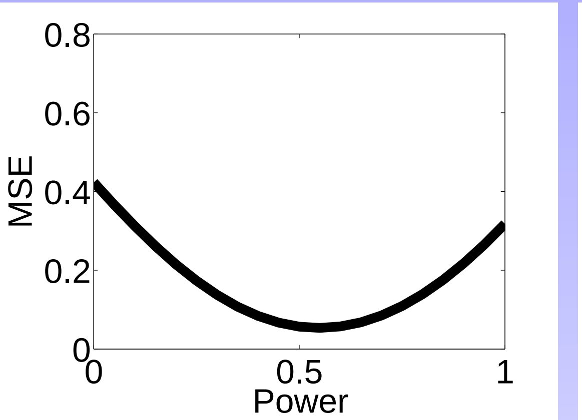
Height

$$\bar{r} = 0.0347$$

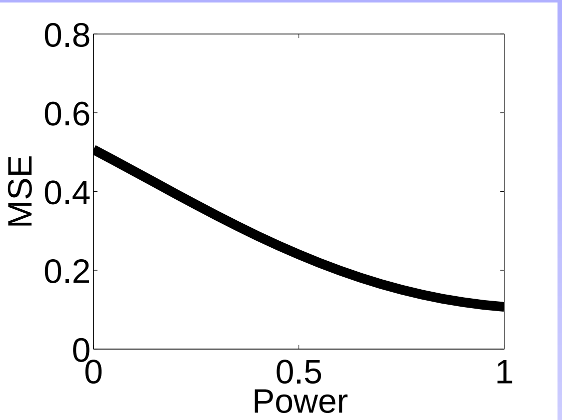
$n = 432$



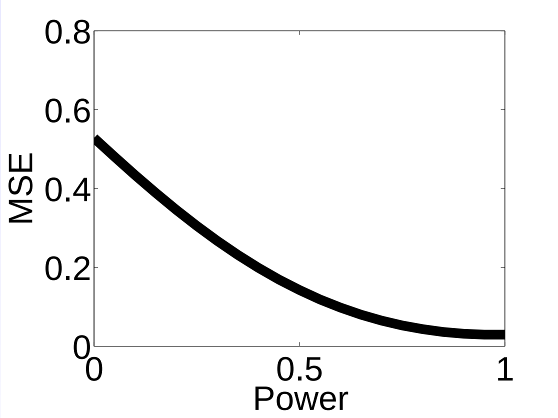
$n = 864$



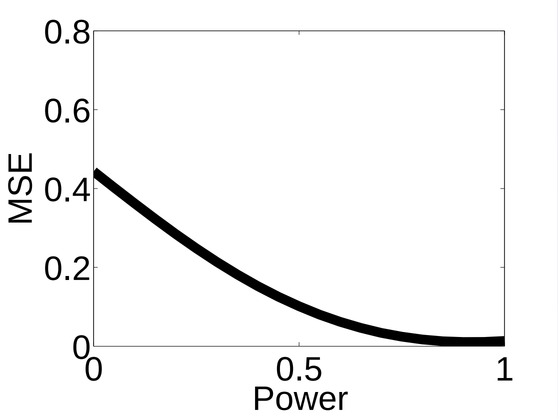
$n = 1730$



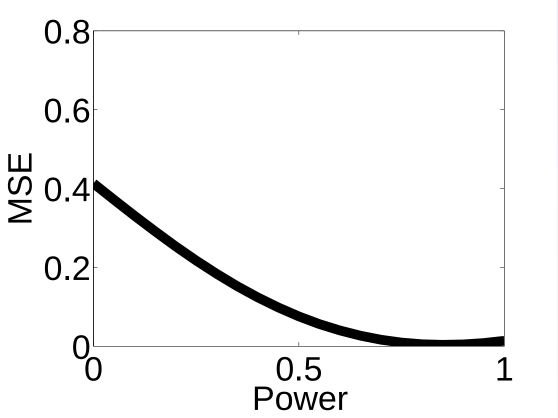
$n = 3460$



$n = 6918$

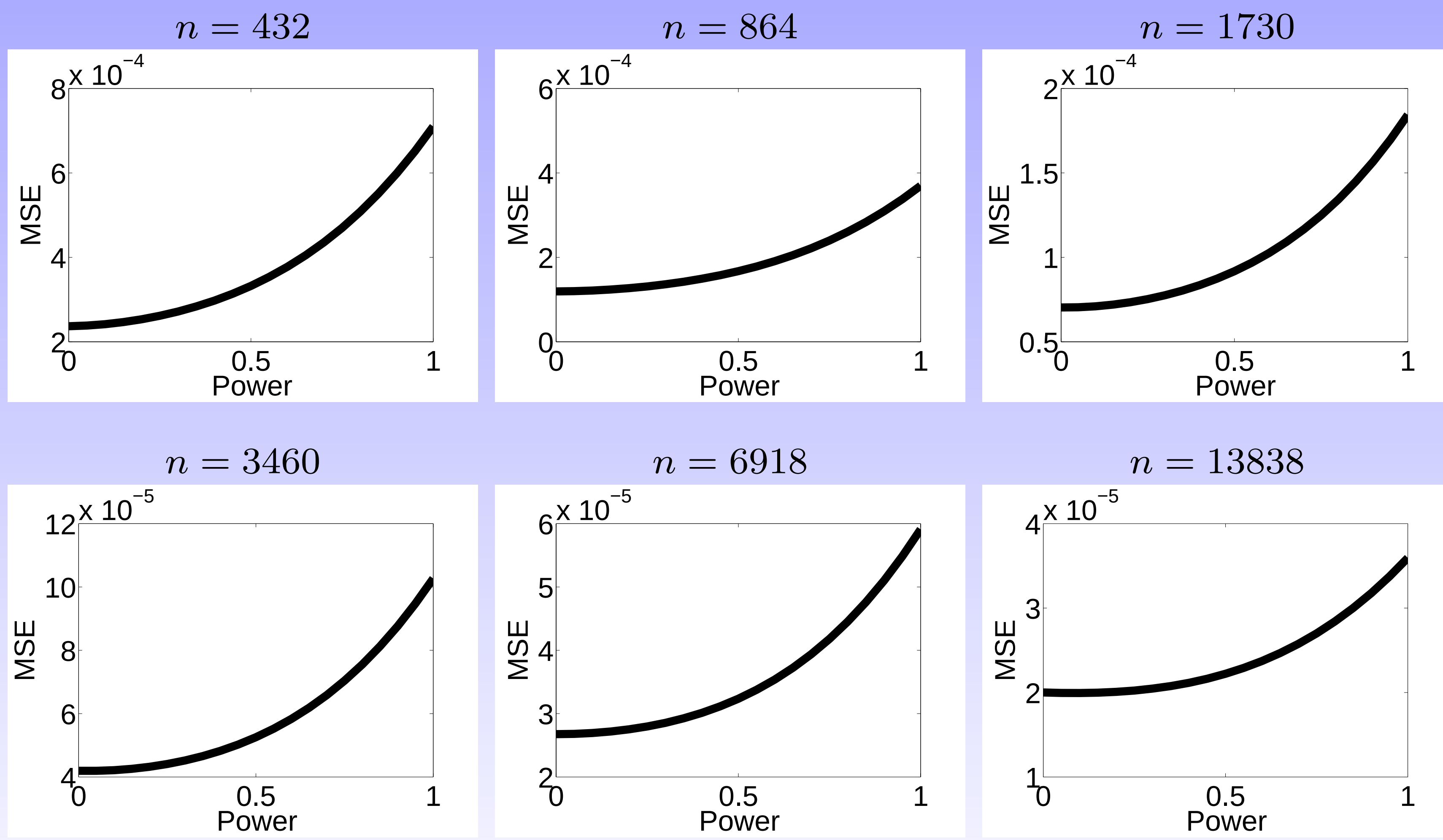


$n = 13838$



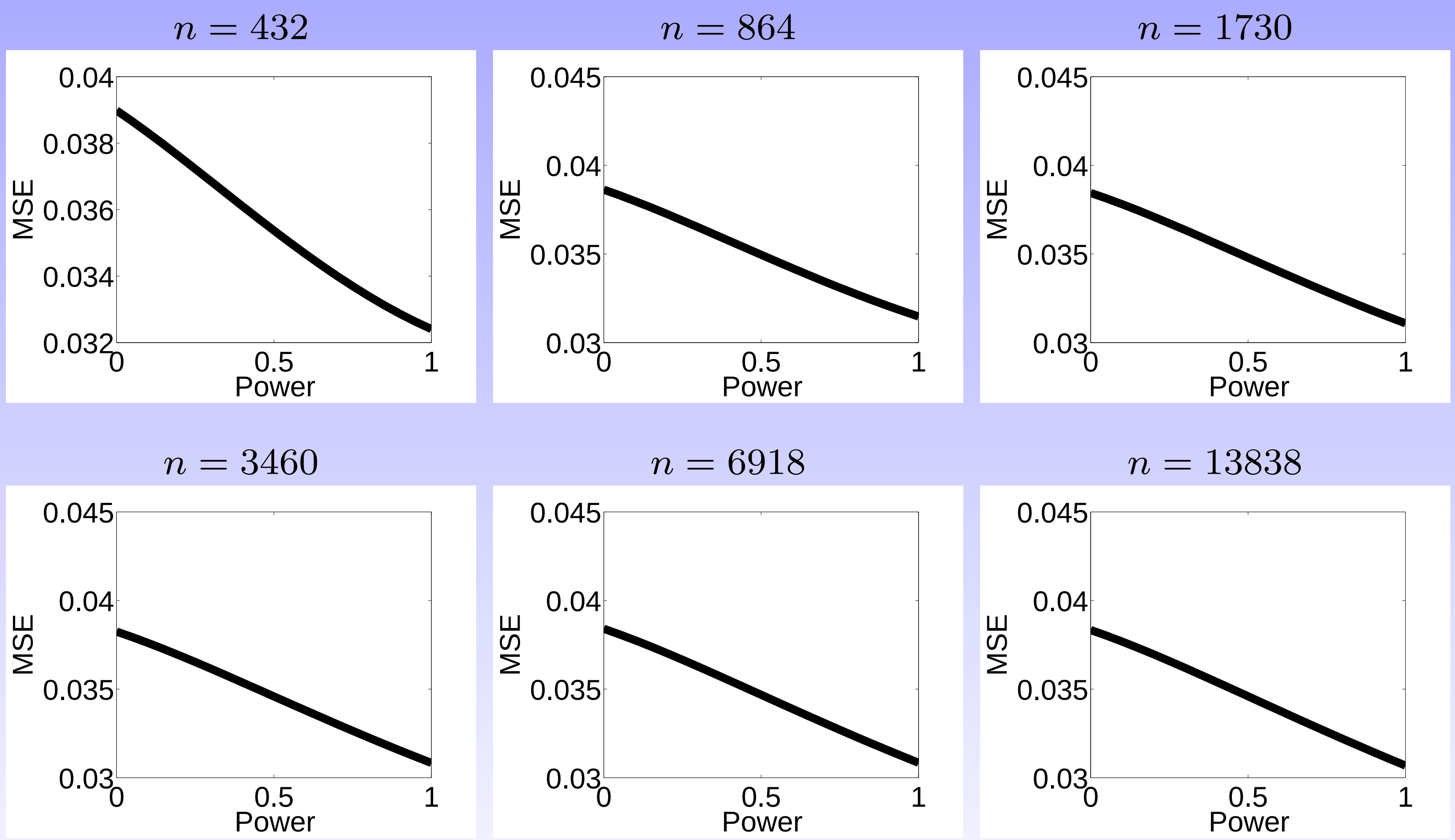
Social Phobia

$$\bar{r} = \text{-}0.0002$$



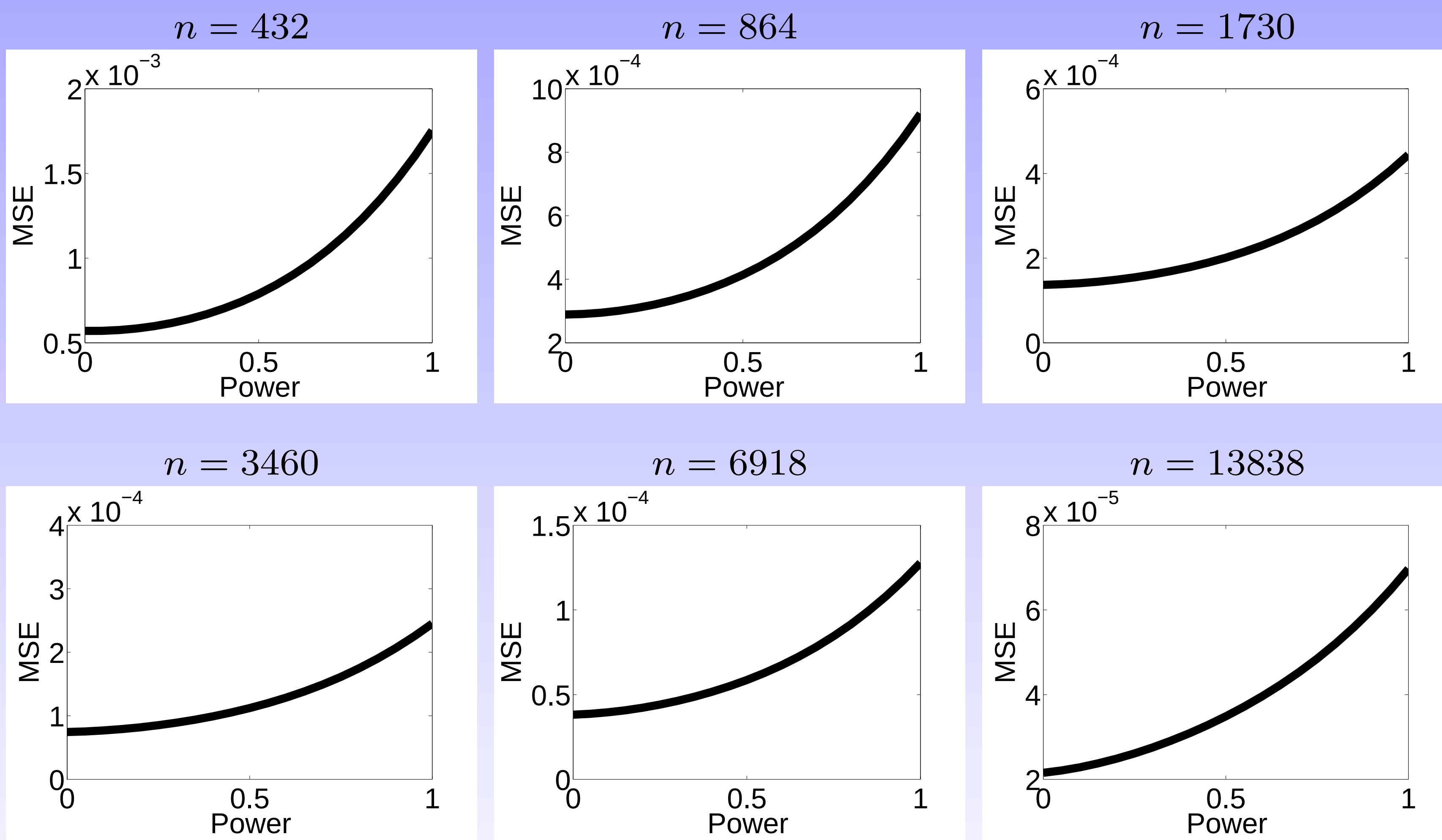
Immigrant

$\bar{r} =$ -0.0296



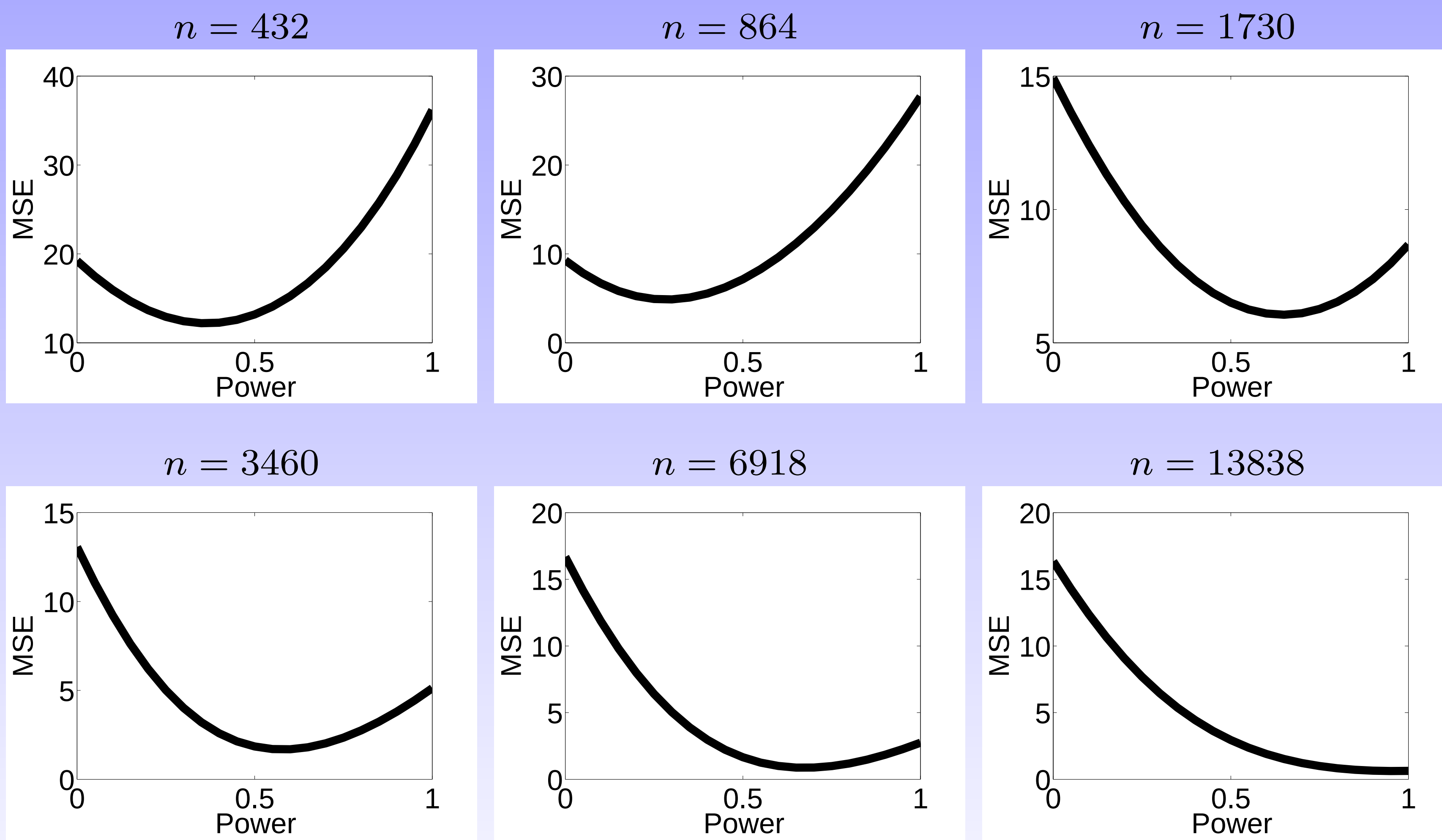
Any Disorder

$$\bar{r} = \text{-}0.0001$$



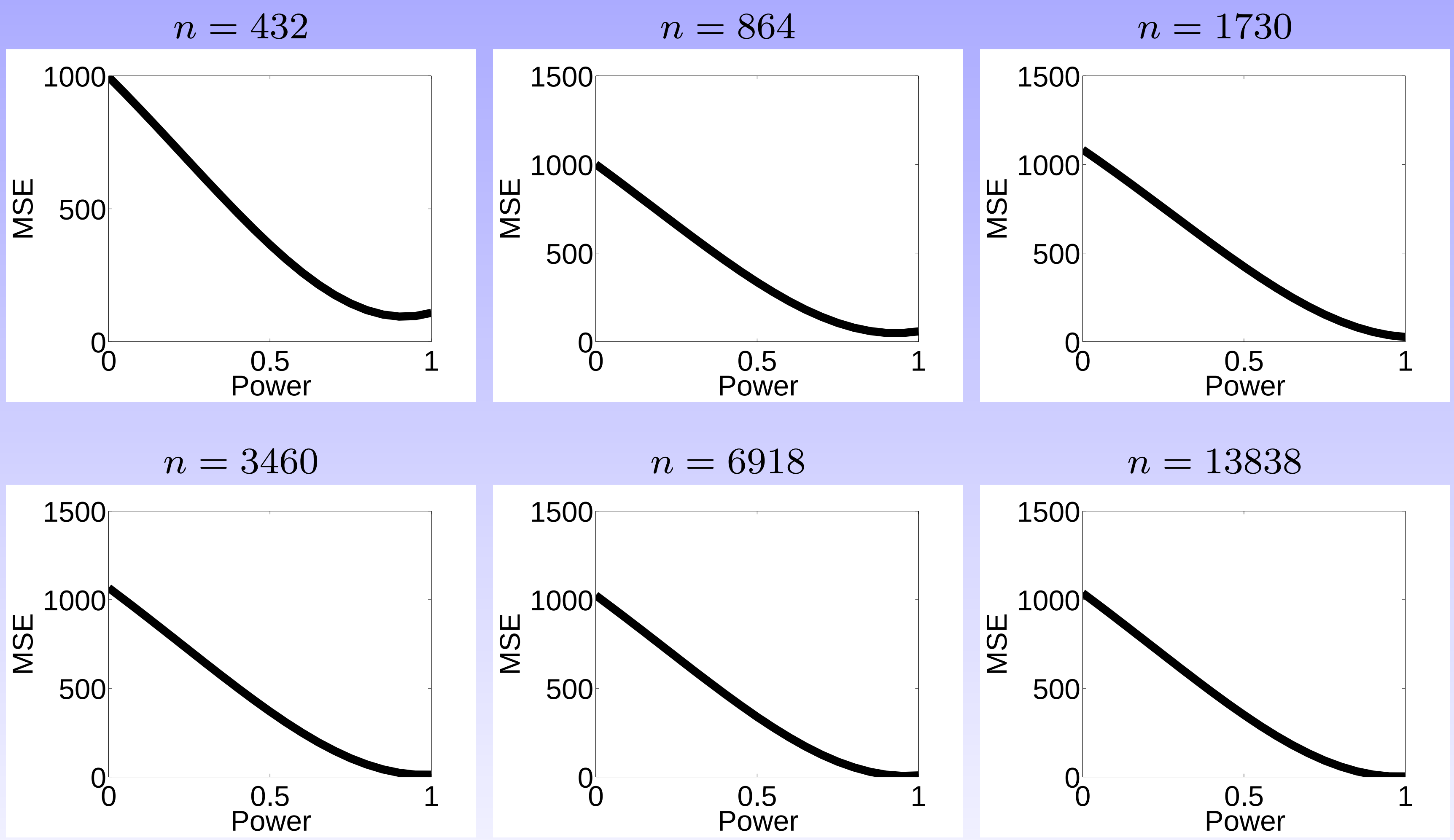
Body Weight

$$\bar{r} = 0.0520$$



Ageplusswgt

$$\bar{r} = 0.5360$$



So What Do We Observe?

√ The optimal bias-variance trade-off depends on both r and n.

So What Do We Observe?

- √ The optimal bias-variance trade-off depends on both r and n.
- √ As n increases, p_{opt} approaches $p = 1$ (as it should!).

So What Do We Observe?

- ✓ The optimal bias-variance trade-off depends on both r and n.
- ✓ As n increases, p_{opt} approaches $p = 1$ (as it should!).
- ✓ But for r very small (e.g., $|r| < 0.01$), a quite large n is needed before we prefer a large power.

So What Do We Observe?

- The optimal bias-variance trade-off depends on both r and n.
- As n increases, p_{opt} approaches $p = 1$ (as it should!).
- But for r very small (e.g., $|r| < 0.01$), a quite large n is needed before we prefer a large power.
- Suggest to predict p_{opt} via $\hat{r}$ and n:

$$\log\left(\frac{p_{opt}}{1 - p_{opt}}\right) = \beta_0 + \beta_1 \log\left(\frac{|\hat{r}|}{1 - |\hat{r}|}\right) + \beta_2 \log(n)$$

So What Do We Observe?

- ✓ The optimal bias-variance trade-off depends on both r and n.
- ✓ As n increases, p_{opt} approaches $p = 1$ (as it should!).
- ✓ But for r very small (e.g., $|r| < 0.01$), a quite large n is needed before we prefer a large power.
- ✓ Suggest to predict p_{opt} via $\hat{r}$ and n:

$$\log\left(\frac{p_{opt}}{1-p_{opt}}\right) = \beta_0+\beta_1 \log\left(\frac{|\hat{r}|}{1-|\hat{r}|}\right)+\beta_2 \log(n)$$

- ✓ Note $\hat{r}$ is the *unweighted* sample correlation between y and w.

Empirical Findings

√ Using the entire CPES data

Variable	Coefficient	t-probability	Confidence Interval	
$Intercept$	-1.3255	0.5097	-5.3079	1.5391
$\mathrm{logit}(\|\hat{\mathrm{r}}\|)$	0.9327	0.0000	0.6338	1.0220
$\log(n)$	0.7421	0.0036	0.3063	1.1548

$R^2 = 0.5446$

Empirical Findings

√ Using the entire CPES data

Variable	Coefficient	t-probability	Confidence Interval	
Intercept	-1.3255	0.5097	-5.3079	1.5391
$\mathrm{logit}(\lvert\hat{r}\rvert)$	0.9327	0.0000	0.6338	1.0220
$\log(n)$	0.7421	0.0036	0.3063	1.1548

$R^2 = 0.5446$

√ Using the NLAAS dataset only:

Variable	Coefficient	t-probability	Confidence Interval	
Intercept	-3.6990	0.0217	-6.8359	-0.5621
$\mathrm{logit}(\lvert\hat{r}\rvert)$	1.0211	0.0000	0.6335	1.4089
$\log(n)$	1.1308	0.0000	0.7245	1.5372

$R^2 = 0.5012$

KISS: Keep it Sophistically Simple

- ✓ To KISS, we suggest (for now!) $\hat{\beta}_0 = -4, \hat{\beta}_1 = 1, \hat{\beta}_2 = 1$, yielding

$$\hat{p}_n = \frac{n|\hat{r}|}{(1-|\hat{r}|)e^4 + n|\hat{r}|}.$$

KISS: Keep it Sophistically Simple

- ✓ To KISS, we suggest (for now!) $\hat{\beta}_0 = -4, \hat{\beta}_1 = 1, \hat{\beta}_2 = 1$, yielding

$$\hat{p}_n = \frac{n|\hat{r}|}{(1 - |\hat{r}|)e^4 + n|\hat{r}|}.$$

- ✓ Desire properties of $\hat{p}_n$:

KISS: Keep it Sophistically Simple

- √ To KISS, we suggest (for now!) $\hat{\beta}_0 = -4, \hat{\beta}_1 = 1, \hat{\beta}_2 = 1$, yielding

$$\hat{p}_n = \frac{n|\hat{r}|}{(1-|\hat{r}|)e^4 + n|\hat{r}|}.$$

- √ Desire properties of $\hat{p}_n$:
  - ⋆ For fixed $\hat{r} \neq 0$, it approaches 1 as $n \to \infty$;

KISS: Keep it Sophistically Simple

- √ To KISS, we suggest (for now!) $\hat{\beta}_0 = -4, \hat{\beta}_1 = 1, \hat{\beta}_2 = 1$, yielding

$$\hat{p}_n = \frac{n|\hat{r}|}{(1-|\hat{r}|)e^4 + n|\hat{r}|}.$$

- √ Desire properties of $\hat{p}_n$:
  - ⋆ For fixed $\hat{r} \neq 0$, it approaches 1 as $n \to \infty$;
  - ⋆ $\bar{y}^{(\hat{p}_n)}$ is *asymptotically equivalent* to $\bar{y}^{(1)}$;

KISS: Keep it Sophistically Simple

- √ To KISS, we suggest (for now!) $\hat{\beta}_0 = -4, \hat{\beta}_1 = 1, \hat{\beta}_2 = 1$, yielding

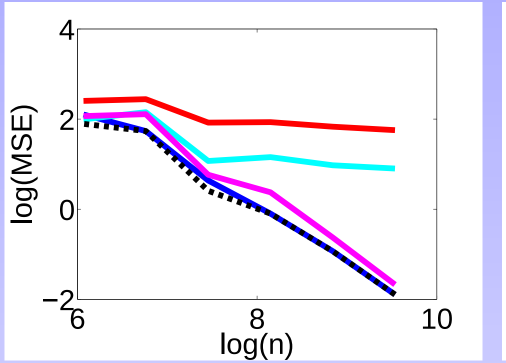
$$\hat{p}_n = \frac{n|\hat{r}|}{(1-|\hat{r}|)e^4 + n|\hat{r}|}.$$

- √ Desire properties of $\hat{p}_n$:
  - ⋆ For fixed $\hat{r} \neq 0$, it approaches 1 as $n \to \infty$;
  - ⋆ $\bar{y}^{(\hat{p}_n)}$ is *asymptotically equivalent* to $\bar{y}^{(1)}$;
  - ⋆ For fixed n, it goes to 1 as $\hat{r} \to 1$ and 0 as $\hat{r} \to 0$.

Log(MSE) under $\{\beta_0 = -4, \beta_1 = 1, \beta_2 = 1\}$.

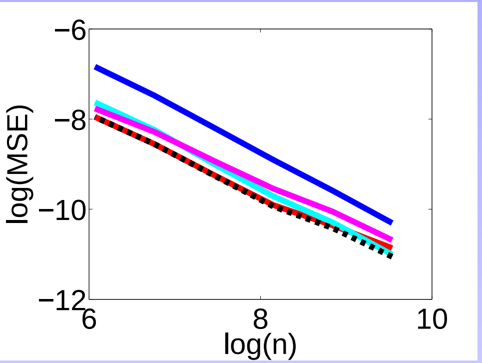
Age

$\bar{r} = 0.0617$



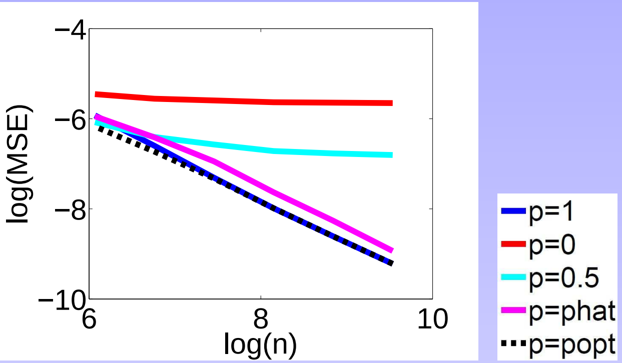
Major Depression

$\bar{r} =$ -0.0069



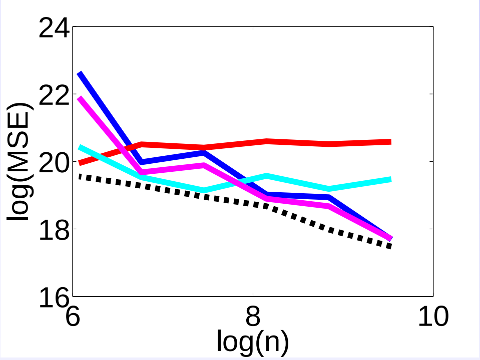
Gender

$\bar{r} =$ -0.0716



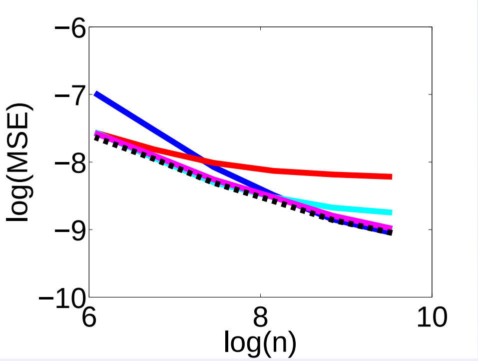
Survey Weight

$\bar{r} = 0.1687$



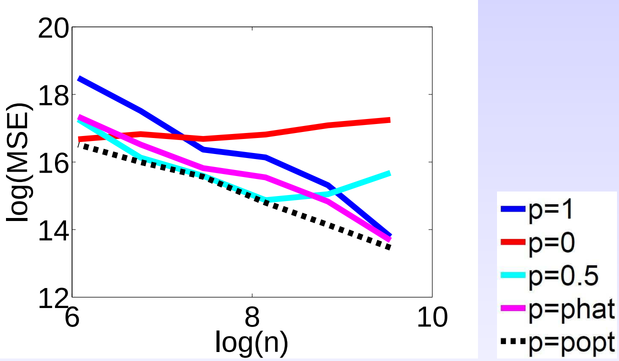
Substance Abuse

$\bar{r} = 0.0139$



Household Income

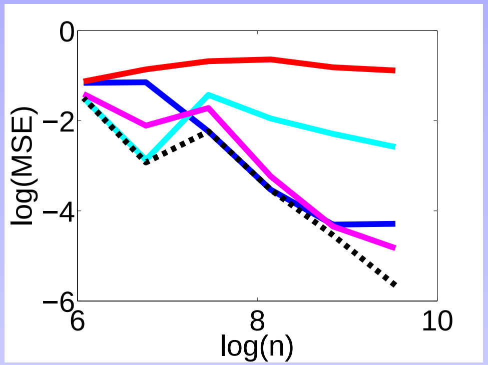
$\bar{r} = 0.0809$



Log(MSE) under $\{\beta_0 = -4, \beta_1 = 1, \beta_2 = 1\}$.

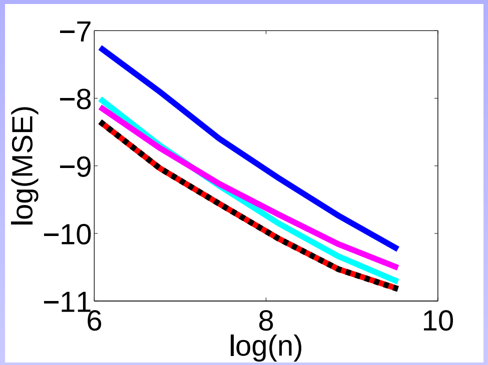
Height

$\bar{r}$ = 0.0347



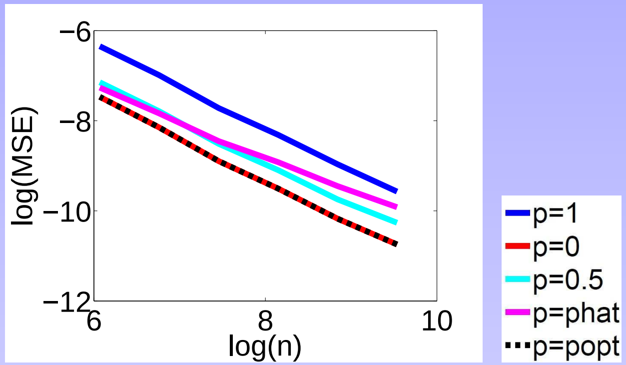
Social Phobia

$\bar{r}$ = -0.0002



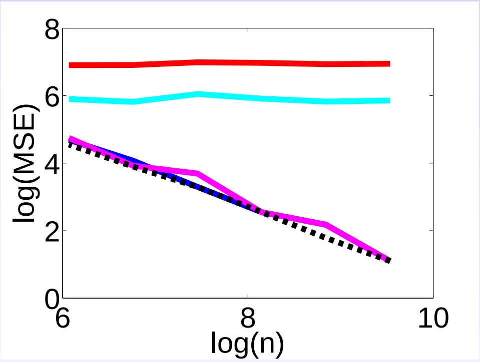
Any Disorder

$\bar{r}$ = -0.0001



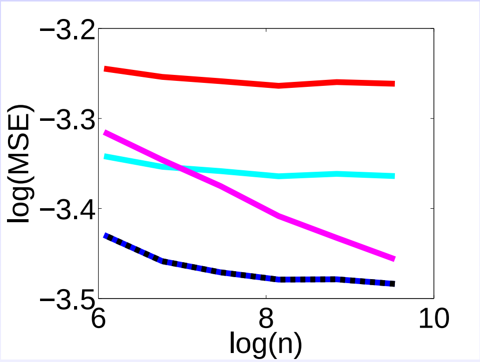
Ageplusweight

$\bar{r}$ = 0.5360



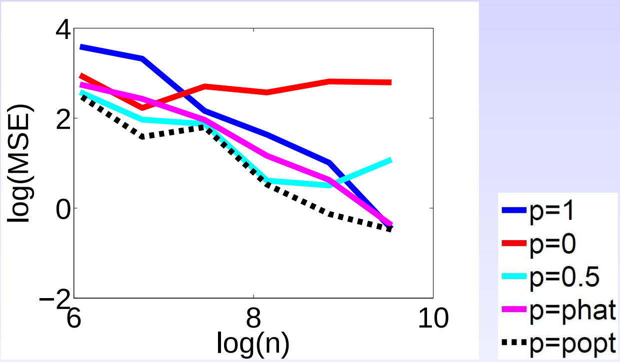
Immigrant

$\bar{r}$ = -0.0296



Body Weight

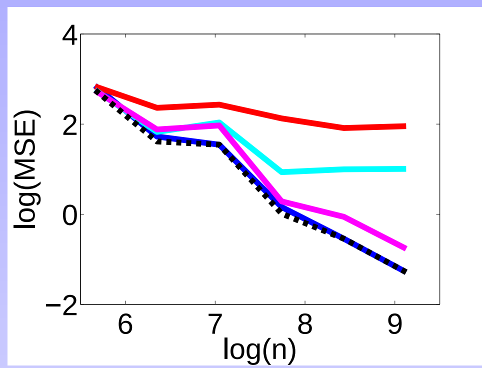
$\bar{r}$ = 0.0520



Out of Sample, Log(MSE) under $\{-4, 1, 1\}$

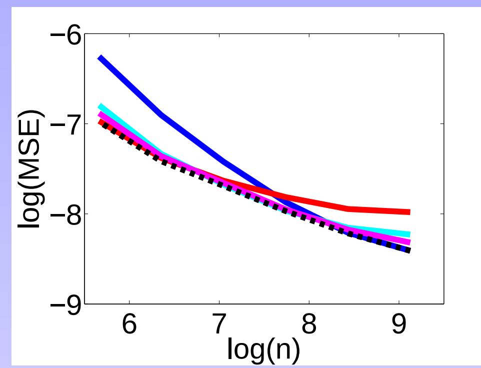
Age

$\bar{r}$ = 0.0259



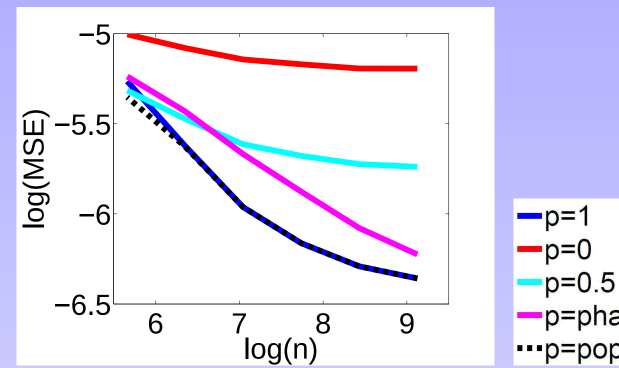
Major Depression

$\bar{r}$ = -0.0075



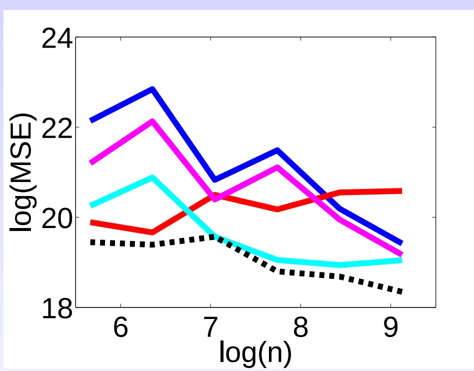
Gender

$\bar{r}$ = -0.0516



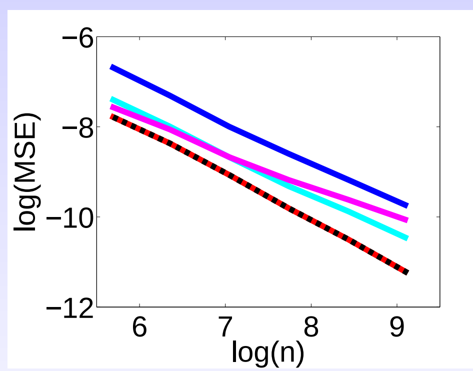
Survey Weight

$\bar{r}$ = 0.1811



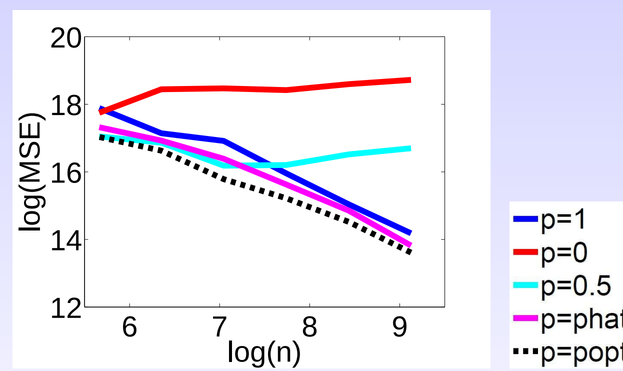
Substance Abuse

$\bar{r}$ = 0.0058



Household Income

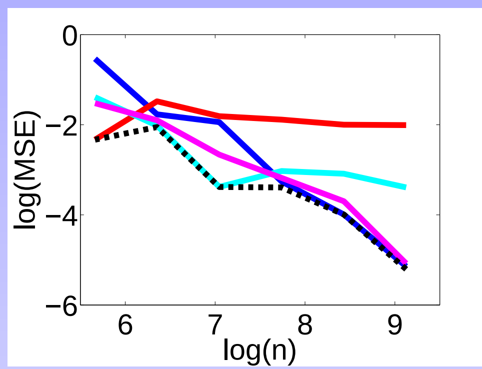
$\bar{r}$ = 0.1050



Out of Sample, Log(MSE) under $\{-4, 1, 1\}$

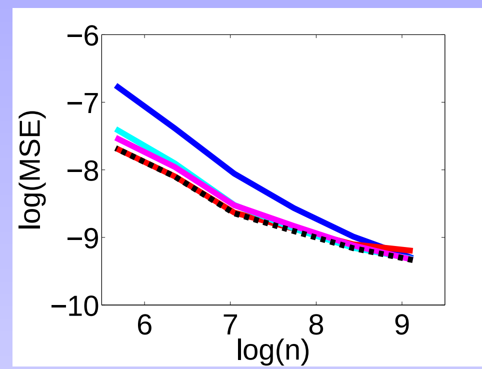
Height

$\bar{r}$ = 0.0673



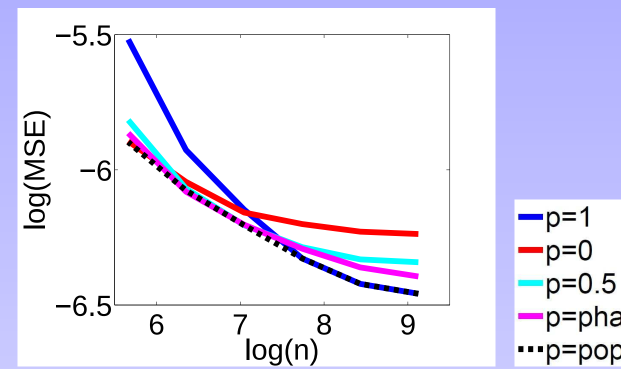
Social Phobia

$\bar{r}$ = -0.0054



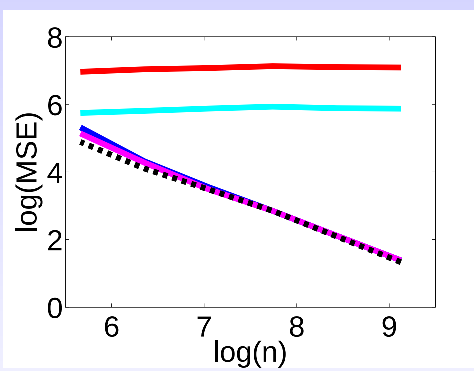
Any Disorder

$\bar{r}$ = -0.0068



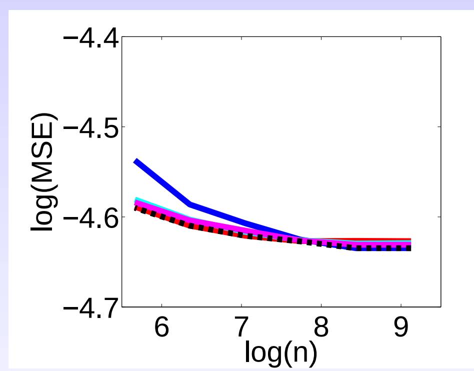
Ageplusswgt

$\bar{r}$ = 0.5219



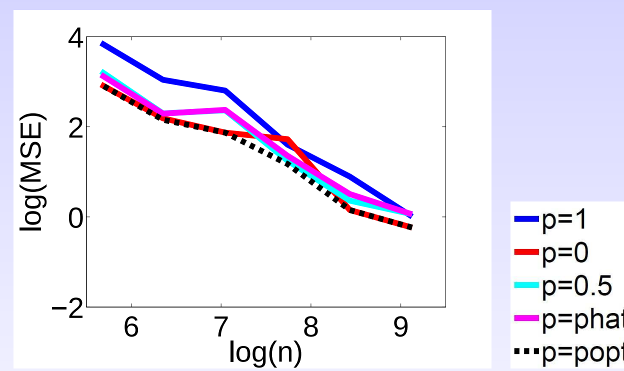
Immigrant

$\bar{r}$ = -0.0020



Body Weight

$\bar{r}$ = -0.0118



Weighting is Still a Mess

√ Using power is a promising general framework.

Weighting is Still a Mess

- √ Using power is a promising general framework.
- √ But a lot more research is needed, both empirical and theoretical.

Weighting is Still a Mess

- √ Using power is a promising general framework.
- √ But a lot more research is needed, both empirical and theoretical.
- √ A common power for all estimators v.s. individual power for specific estimators.

Weighting is Still a Mess

- √ Using power is a promising general framework.
- √ But a lot more research is needed, both empirical and theoretical.
- √ A common power for all estimators v.s. individual power for specific estimators.
- √ How to find “nearly optimal" power for weighted regressions, especially binary regressions?

Weighting is Still a Mess

- √ Using power is a promising general framework.
- √ But a lot more research is needed, both empirical and theoretical.
- √ A common power for all estimators v.s. individual power for specific estimators.
- √ How to find “nearly optimal" power for weighted regressions, especially binary regressions?
- √ Is it possible to find an adaptively optimal power?

Weighting is Still a Mess

- √ Using power is a promising general framework.
- √ But a lot more research is needed, both empirical and theoretical.
- √ A common power for all estimators v.s. individual power for specific estimators.
- √ How to find “nearly optimal" power for weighted regressions, especially binary regressions?
- √ Is it possible to find an adaptively optimal power?
- √ ...