Power-Shrinkage and Trimming: Two Ways to Mitigate Excessive Weights

Chihnan Chen[1], Nanhua Duan[2], Xiao-Li Meng[3], Margarita Alegria[4]
Boston University[1]
UCLA[2]
Harvard University[3]
Cambridge Health Alliance and Harvard Medical School[4]

Abstract

Large-scale surveys often produce raw weights with very large variations. A standard approach is to perform some form of trimming, as a way to reduce potentially large variances in various survey estimators. The amount of trimming is usually determined by considerations of bias-variance trade-off. While bias-variance trade-off is a sound principle, the trimming method itself is popular only because of its simplicity, not because it has statistically desirable properties. In this paper we investigate a more principled method by shrinking the variability of the log of the weights. We use the same mean squared error (MSE) criterion to determine the amount of shrinking. The shrinking on the log scale implies shrinking weights by a power parameter p between [0,1]. Our investigation suggests an empirical way to predict the optimal choice. This power shrinkage method provides a natural way to deal with measurement errors and outliers in the raw weights, while preserving the ranking of the raw weights. We demonstrate the use of this method with the National Latino and Asian American Study (NLAAS).

Keywords: Selection probability, unequal selections, selection biases, self-weighting.

1 Introduction

In survey studies, the weights are often used to mitigate the bias from the unequal selection probability. Estimation bias is reduced by sampling weight at the expense of increased variance. As an example, Kish (1992) provided an authoritative account of the weighted and unweighted estimators. He discussed the weight trimming approach to reduce the mean square error through the bias-variance trade-off. We propose below a new approach, power-shrinkage method, to deal with this problem. The power-shrinkage approach we propose is a continuous transformation and preserves the ranking of weights. We investigate the properties of both trimming and power-shrinkage approaches via simulation studies.

We organize this paper in five sections. Section 1 is the introduction. Section 2 provides the formulas of shrinkage weights. Section 3 describes the NLAAS data set used in our simulation study. Section 4 describes the simulation design. Section 5 summarizes the results and concludes.

2 Power-Shrinkage and Trimming

The survey weights are constructed as the reciprocals of the selection probabilities. Let w be the survey weight, n be the sample size, and y be the variable of interest. Let $p \in [0, 1]$ be the power-shrinkage parameter and $T \in [0, 1]$ be the trimming threshold defined in terms of percentile. The original weighted estimator, power-shrinkage estimator and the trimmed estimator for the expectation, $E(y)$, are given by

- Original weighted estimator : $\bar{y}_w = \dfrac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$
- Power-Shrinkage estimator: $\bar{y}_w^{(p)} = \dfrac{\sum_{i=1}^{n} w_i^p y_i}{\sum_{i=1}^{n} w_i^p}$
- Trimmed weighted estimator: $\bar{y}_w^{(T)} = \dfrac{\sum_{i=1}^{n} w_i(T) y_i}{\sum_{i=1}^{n} w_i(T)}$

where $w_i(T) = \min(w_i, Z(T))$, and $Z(T)$ is the $(100 \times T^{th})$ percentile of w.

It is well known that while $\bar{y}_w$ is unbiased, its variance can be very large when the variance of the weights is large. The trimming estimate $\bar{y}_w^{(T)}$ addresses this problem by reducing those excessively large weights via $w_i(T) = \min(w_i, Z(T))$, that is, the top weights above $Z(T)$ are trimmed down to $Z(T)$. This trimming will introduce bias, but the hope is that the variance of $\bar{y}_w^{(T)}$ is much smaller than $\bar{y}_w$, such that the overall mean square error of $\bar{y}_w^{(T)}$ is smaller than $\bar{y}_w$.

Although trimming is a very useful and popular method, it is nevertheless an ad-hoc method, and has a number of undesirable properties. For example, it does not preserve the (strict) ranks of the original weights. Furthermore, it does not address problems in the construction of the weights that are likely to affect all or most weights even though most of them did not become excessive.

We therefore propose the power-shrinkage method based on the motivation described below. In most survey studies, the distribution of the weights is reasonably approximated by the log normal distribution (see section 3 for a real data example). So if we want to reduce the impact of excessive weights, or if we believe that there is noise in the weights (e.g. due to measurement error in the construction of weights), we should rescale the log of the weights to reduce its variance. This is the same as raising the weight by a power $p \in [0, 1]$, which gives the power-shrinkage estimate $\bar{y}_w^{(p)}$.

When $p = 1$ and $T = 1$, there is no shrinkage or trimming; all three estimators are the same; $\bar{y}_w = \bar{y}_w^{(p)} = \bar{y}_w^{(T)}$. On the other hand, when $p = 0$ and $T = 0$, both power-shrinkage and trimmed estimators are equivalent to the unweighted estimator. When p and T are close to one, the estimators are less biased but with larger variances. When p and T are close to zero, they behave like the unweighted estimator, biased but with a smaller variance. The optimal choice of p depends on a number of factors. The challenge is therefore to find a practical way to find a good (not necessary optimal) choice of p, as well as some common choices of p that will work reasonably well in a variety of situations. To explore these and to compare power-shrinkage and trimming approaches, we designed a simulation study using the NLAAS dataset as the template.

3 NLAAS

The National Latino and Asian American Study (NLAAS) is a nationally representative survey of White, Latino and Asian American household residents (aged 18 and older) in the non-institutionalized population of the US. A total of 4864 individuals, including Latinos, Asians, and Whites, were interviewed. The sample includes an NLAAS Core, designed to be nationally representative of all Latino origin groups regardless of geographic patterns; and NLAAS-HD supplements, designed to oversample geographic areas with moderate to high density (HD) of Latino and Asian households. Weighting reflecting the joint probability of selection from the pooled Core and HD samples provides sample-based coverage of the national Latino population.

The properties of weights in NLAAS are summarized in Table 1, Figure 1 and Figure 2. As we can see, the variance of the weight is large, and the distribution of log weight is roughly Gaussian.

	w	Percentile	w
MIN	80	5	368
MEAN	7,340	25	1,325
MAX	136,011	50	2,861
		75	8,625
		95	28,385

Table 1: Summery Statistics for w

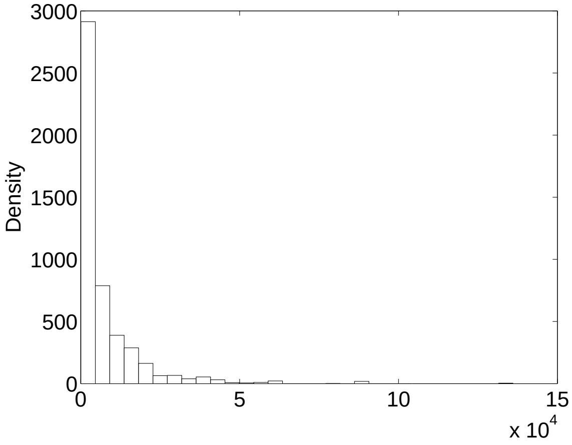

Figure 1: Distribution of Survey Weight w

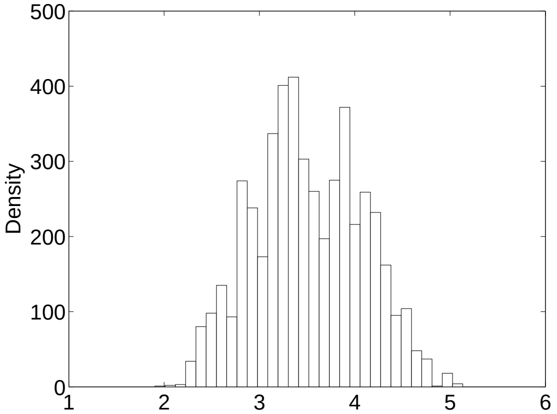

Figure 2: Distribution of $\log_{10}(w)$

4 Simulation Design

To investigate the properties of power-shrinkage and trimming approaches, we designed a simulation study using the NLAAS dataset. We constructed a hypothetical population consisting of I clusters, where $I = 4,864$. Assume there are $j = 1, ..., w_i$ individuals in the cluster i, and their $y_{i,j} \sim N(y_i, \sigma_y^2)$, where σ_y is the standard deviation of y in each cluster, which we use the unweighted standard error from the y in the NLAAS dataset. Our artificial population consisted of $N = \sum_{i=1}^{I} w_i = 35,705,416$ individuals. If y is restricted to be positive, we use the log-normal distribution instead of the normal distribution. If y is a binary variable other than gender, we use a logistic regression on age and gender to predict the outcomes in each cluster. The variable gender is predicted with a logistic regression model on age.

We applied a two-stage sampling design. First, we draw q clusters by simple random sampling without replacement. Second, we 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 $w_i^* = w_i/s$.

Given the sample design, we are interested in the following estimators:

- The sample estimator:

$$\bar{y}_w = \frac{\sum_{i=1}^{q}\sum_{j=1}^{s} w_i^* y_{i,j}}{\sum_{i=1}^{q}\sum_{j=1}^{s} w_i^*}$$

- The Power-Shrinkage estimator:

$$\bar{y}_w^{(p)} = \frac{\sum_{i=1}^{q}\sum_{j=1}^{s} (w_i^*)^p y_{i,j}}{\sum_{i=1}^{q}\sum_{j=1}^{s} (w_i^*)^p}$$

- The trimmed estimator:

$$\bar{y}_w^{(T)} = \frac{\sum_{i=1}^{q}\sum_{j=1}^{s} w_i^*(T) y_{i,j}}{\sum_{i=1}^{q}\sum_{j=1}^{s} w_i^*(T)}$$

In each iteration, the power-shrinkage and trimmed estimators with shrinkage parameters/trimmed thresholds at $0, 0.05, 0.1, 0.15, ..., 0.95, 1$ are calculated. The optimal shrinkage parameter is denoted as p^*, and the optimal trimmed threshold is T^*. To compare the minimal mean square error of power-shrinkage approach against that of the trimming approach, we compute the ratio of the optimal power-shrinkage MSE (PMSE) to the optimal trimming MSE (TMSE):

$$R_{pt} = \frac{PMSE(p^*)}{TMSE(T^*)}.$$

We also explore the properties of both approaches given different sample sizes. We set $s = 2$ and change the value of q to control the sample size. In this simulation, we use $q = 2432, 1216, 608, 304, 152, 76$. So the sample size is $n = qs = 4864, 2432, 1216, 608, 304, 152$, respectively.

It is well known that the correlation between the weights and the variable to be analyzed plays an important role in determining the bias of the unweighted estimator. If the correlation is low, then the bias of the unweighted estimator is small. We therefore selected variables with different correlations with the weights to study this issue.

Specifically, we chose the following variables: gender, age, height, education, major depression, substance disorder, social phobia, any disorder, immigrant status, k10-distress measure, agepluswgt, and survey weight w. Here the variable agepluswgt is constructed as

$$\text{agepluswgt} = \text{age} + 0.001 \times w,$$

which has a correlation of 0.5726 with w. By investigating agepluswgt and survey weight w, we attempt to learn the performance of power-shrinkage and trimming for variables with high correlation with survey weight. Note that the correlation between variables and weights in NLAAS, ρ, is different from the correlation in our artificial population, which is denoted as $\hat{\rho}$, typically lower in magnitude because the random noise we introduced within each cluster.

To investigate the relationship between the optimal power-shrinkage parameter p^*, trimming threshold T^*, simulated correlation $\hat{\rho}$ and sample size n, we fit the following regression model.

$$\log\left(\frac{p^*}{1-p^*}\right) = \beta_0 + \beta_1 \log\left(\frac{|\hat{\rho}|}{1-|\hat{\rho}|}\right) + \beta_2 \log(n) + \varepsilon \quad (1)$$

$$\log\left(\frac{T^*}{1-T^*}\right) = \gamma_0 + \gamma_1 \log\left(\frac{|\hat{\rho}|}{1-|\hat{\rho}|}\right) + \gamma_2 \log(n) + \varepsilon \quad (2)$$

where the p^* or T^* are replaced as 0.99 if it is 1, and replaced as 0.01 if it is 0. We also fit the regression model with an interaction term, which turns out to be far from significant.

5 Results

We summarize the simulation study results in Table 2 and Figures 3 to 14. Table 2 presents the optimal power-shrinkage parameters p^*, the optimal trimming thresholds T^* and the ratio of optimal mean square errors R_{pt} along with the sample size n and the simulated sample correlation, $\hat{\rho}$, between variable, y, and weights, w^*. In Figure 3 to 14, the mean square error of the power-shrinkage approach is denoted as the solid curve. The mean square error of the trimming approach is denoted as the dotted curve. The horizontal axis indexes the power-shrinkage parameters or trimming thresholds. The vertical axis is the mean square error. The sample size is noted on the top of each graph.

The mean square error curves are generally smooth for the power-shrinkage estimators, but are usually jagged for the trimmed estimators. This is because power-shrinkage is a continuous transformation while trimming is not a smooth operator. But overall, the two methods are comparable in terms of minimizing the MSE because the ratio R_{pt} is quite close to one for most cases listed in Table 2.

Our limited evidences suggest that the power-shrinkage method works well with smaller sample sizes and/or when the correlations between the variable and the weight is high. As expected, we also observe that the optimal bias-variance trade-off varies with the correlation between each variable and the survey weight. When the correlation is high, little trimming or shrinkage is preferred. Our simulation results also suggest that the usual practice of trimming a small fraction of the most extreme weights might not be the best way to trim – it is often better to trim more aggressively or shrink more aggressively, unless the correlation is very high.

Because the survey weight and ageplusweight are not typical variables of interests in practice, we exclude them in our attempt to fit the regression models (1) and (2). For model (1), the least-square estimates for the regression parameters are $\hat{\beta}_0 = -3.70$, $\hat{\beta}_1 = 1.02$ (significant) and $\hat{\beta}_2 = 1.13$ (significant). This result is expected because intuitively the optimal power should increase with both the correlation and the sample size. The reason is with either large correlation or large sample size, the bias in an inapproiately weighted estimator becomes more dominated. Same is true with the trimming, that is, the optimal trimming threshold should increase with the correlation or sample size. For model (2), the regression estimates are $\hat{\gamma}_0 = -3.19$, $\hat{\gamma}_1 = 1.24$ (significant) and $\hat{\gamma}_2 = 1.27$ (significant). The R^2 for both regressions are about 50%, indicating the usefulness of these two simple models.

The results above inspired us to seek a simple approximation formula for predicting the optimal shrinking power based on the correlation and sample size. We seek a simple approximation because the regression results here are based on only 10 specifically chosen variables, so while the results are suggestive, they should not be taken literally for general consumption. Instead, because $\hat{\beta}_0 = -3.70, \hat{\beta}_1 = 1.02, \hat{\beta}_2 = 1.13$, we suggest that a simple rule to calculate the optimal power-shrinkage parameter is to use $\beta_0^* = -4, \beta_1^* = 1, \beta_2^* = 1$, that is

$$\hat{p} = \frac{n|\hat{\rho}|e^{-4}}{1 - |\hat{\rho}| + n|\hat{\rho}|e^{-4}} \tag{3}$$

The resulting predictions, rounding to the closest grid point in our simulation study, are reported in Table 3, where we also calculate the ratio $\frac{MSE(\hat{p})}{MSE(p^*)}$ to evaluate how this simple rule performs. The results are very encouraging, because the loss of optimality by this approximation is negligible in most cases in terms of the MSE. Even for the two out-sample predictions – recall survey weights and ageplusweight were not used in the regression – the results are not bad, especially for ageplusweight. Of course, much more investigations are needed, both theoretical and empirical, to test to what extend the simple rule (3) is useful. Our conjecture is that the use of unit slope for both predicting variables should hold fairly generally, but the value of the intercept may need to be changed non-trivially for some other variables.

While simple rules such as (3) are quite useful for constructing efficient estimators for individual variables, they are not useful when we want to seek a single power to construct a fixed set of weights to be used for any variables measured in a survey. For the latter purpose, we could seek a "minmax" type power, that is, a power shrinkage that would minimize the worst MSE across all variables of interests. Our simulation results show that across the 12 variables in our simulation studies, the minimax choice of p is in the range of $[0.3, 0.8]$ across different sample sizes. Because one can easily construct variables to be completely correlated or uncorrelated with the weights, we believe that the minimax choice of p over enough choices of variables will converge to $p = 0.5$. Further theoretical investigation is planned to validate this conjuncture.

References

Alegria, M., Takeuchi, D., Canino, G., Duan, N., Shrout, P., Meng, X.L., Vega, W., Zane, N., Vila, D., Woo, M., Vera, M., Guarnaccia, P., Aguilar-Gaxiola, S., Sue, S., Escobar, J., Lin, K-M, Gong, F. (2004). "Considering Context, Place, and Culture: The National Latino and Asian American". *International Journal of Methods in Psychiatric Research*, **13**, 208-220.

Kish, Leslie. (1992). "Weighting for Unequal P_i," *Biometrics*, **8**, 183-200.

Kish, Leslie. (1995). *Survey Sampling*, Wiley Classics Library.

Heeringa, S., Wagner, J., Torres, M., Duan, N., Adams, T., and Berglund, P. (2004). "Sample designs and sampling methods for the Collaborative Psychiatric Epidemiology Studies (CPES)". *International Journal of Methods in Psychiatric Research*, **13**, 221-240.

y	n	$\hat{\rho}$	p^*	T^*	R_{pt}	y	n	$\hat{\rho}$	p^*	T^*	R_{pt}
Age	152	-0.04	0.00	0.00	1.00	Height	152	0.09	0.35	0.60	0.90
$\rho =$	304	-0.07	0.05	0.00	0.99	$\rho =$	304	0.11	0.15	0.20	0.95
-0.0506	608	-0.04	0.15	0.00	0.90	0.0987	608	0.06	0.75	0.95	0.98
	1216	-0.03	0.25	0.45	0.96		1216	0.09	0.75	0.95	0.96
	2432	-0.03	0.45	0.60	1.11		2432	0.08	0.90	0.95	0.89
	4864	-0.03	0.70	0.70	1.93		4864	0.07	0.95	1.00	0.90
Major	152	0.00	0.00	0.00	1.00	Social	152	0.00	0.00	0.00	1.00
Depression	304	0.00	0.00	0.00	1.00	Phobia	304	0.00	0.05	0.00	1.00
$\rho =$	608	0.00	0.00	0.00	1.00	$\rho =$	608	0.00	0.10	0.05	0.99
-0.0017	1216	0.00	0.00	0.00	1.00	0.0306	1216	0.00	0.20	0.45	0.99
	2432	0.00	0.00	0.00	1.00		2432	0.00	0.35	0.65	1.03
	4864	0.00	0.00	0.00	1.00		4864	0.00	0.50	0.75	1.04
Substance	152	0.02	0.20	0.40	0.98	Immigrant	152	-0.01	0.40	0.75	0.99
Abuse	304	0.02	0.40	0.70	0.98		304	-0.01	0.70	0.90	1.00
$\rho =$	608	0.02	0.65	0.90	1.01	$\rho =$	608	-0.02	1.00	1.00	1.00
0.0705	1216	0.02	0.90	0.95	1.02	-0.1843	1216	-0.01	1.00	1.00	1.00
	2432	0.02	1.00	1.00	1.00		2432	-0.01	1.00	1.00	1.00
	4864	0.02	1.00	1.00	1.00		4864	-0.01	1.00	1.00	1.00
Gender	152	-0.07	0.35	0.30	0.93	Any	152	0.01	0.15	0.20	0.99
$\rho =$	304	-0.07	0.45	0.75	0.89	Disorder	304	0.01	0.25	0.45	0.99
-0.0584	608	-0.07	0.65	0.90	0.93	$\rho =$	608	0.01	0.40	0.70	1.03
	1216	-0.07	0.75	0.95	0.94	0.0452	1216	0.01	0.55	0.80	1.07
	2432	-0.07	0.80	0.95	0.92		2432	0.01	0.75	0.90	1.06
	4864	-0.07	0.85	0.95	0.77		4864	0.01	0.90	0.95	1.06
Education	152	-0.08	0.50	0.75	1.24	K10-	152	-0.02	0.05	0.00	1.00
$\rho =$	304	-0.08	0.65	0.80	1.44	Distress	304	-0.03	0.50	0.90	0.93
-0.0883	608	-0.07	0.60	0.70	1.57	$\rho =$	608	-0.03	0.60	0.90	0.94
	1216	-0.05	0.75	0.80	1.70	-0.0333	1216	-0.03	0.80	0.95	1.02
	2432	-0.06	0.85	0.95	2.07		2432	-0.04	0.85	0.95	1.05
	4864	-0.06	0.85	0.85	3.14		4864	-0.02	0.90	1.00	0.96
Survey	152	0.19	0.05	0.50	1.17	Ageplus-	152	0.38	0.70	0.95	1.16
Weight	304	0.24	0.25	0.65	1.48	weight	304	0.36	0.80	0.95	1.07
$\rho =$	608	0.19	0.35	0.80	1.68	$\rho =$	608	0.35	0.90	1.00	0.79
1.0000	1216	0.19	0.50	0.90	1.82	0.5726	1216	0.40	0.90	1.00	0.69
	2432	0.19	0.80	0.95	1.40		2432	0.38	1.00	1.00	1.00
	4864	0.19	0.75	0.95	1.48		4864	0.39	0.95	1.00	0.77

Table 2: Simulation Study Results

y	n	$\hat{p}$	p^*	$\frac{MSE(\hat{p})}{MSE(p^*)}$	y	n	$\hat{p}$	p^*	$\frac{MSE(\hat{p})}{MSE(p^*)}$
Age	152	0.10	0.00	1.06	Height	152	0.25	0.35	1.05
	304	0.30	0.05	1.50		304	0.40	0.15	1.83
	608	0.30	0.15	1.09		608	0.40	0.75	1.40
	1216	0.45	0.25	1.45		1216	0.65	0.75	1.08
	2432	0.60	0.45	1.26		2432	0.80	0.90	1.08
	4864	0.70	0.70	1.00		4864	0.85	0.95	1.30
Major Depression	152	0.00	0.00	1.00	Social Phobia	152	0.00	0.00	1.00
	304	0.00	0.00	1.00		304	0.00	0.05	1.00
	608	0.05	0.00	1.01		608	0.00	0.10	1.01
	1216	0.05	0.00	1.01		1216	0.10	0.20	1.01
	2432	0.10	0.00	1.04		2432	0.10	0.35	1.03
	4864	0.15	0.00	1.10		4864	0.25	0.50	1.03
Substance Abuse	152	0.05	0.20	1.02	Immigrant	152	0.05	0.40	1.02
	304	0.10	0.40	1.07		304	0.05	0.70	1.05
	608	0.20	0.65	1.12		608	0.15	1.00	1.09
	1216	0.35	0.90	1.20		1216	0.25	1.00	1.10
	2432	0.50	1.00	1.25		2432	0.40	1.00	1.09
	4864	0.65	1.00	1.23		4864	0.55	1.00	1.07
Gender	152	0.15	0.35	1.05	Any Disorder	152	0.05	0.15	1.01
	304	0.30	0.45	1.06		304	0.05	0.25	1.03
	608	0.45	0.65	1.10		608	0.10	0.40	1.07
	1216	0.60	0.75	1.11		1216	0.20	0.55	1.10
	2432	0.75	0.80	1.04		2432	0.35	0.75	1.11
	4864	0.85	0.85	1.00		4864	0.50	0.90	1.10
Education	152	0.20	0.50	1.24	K10 Distress	152	0.05	0.05	1.00
	304	0.30	0.65	1.38		304	0.15	0.50	1.06
	608	0.45	0.60	1.11		608	0.30	0.60	1.14
	1216	0.55	0.75	1.15		1216	0.40	0.80	1.35
	2432	0.75	0.85	1.03		2432	0.60	0.85	1.34
	4864	0.85	0.85	1.00		4864	0.65	0.90	1.46
Survey Weight	152	0.40	0.05	2.95	Agepluswgt	152	0.65	0.70	1.01
	304	0.65	0.25	5.05		304	0.75	0.80	1.02
	608	0.75	0.35	4.14		608	0.85	0.90	1.01
	1216	0.85	0.50	3.18		1216	0.95	0.90	1.12
	2432	0.90	0.80	1.16		2432	0.95	1.00	1.05
	4864	0.95	0.75	2.13		4864	1.00	0.95	1.30

Table 3: Checking the performance of the approximated rule (3)

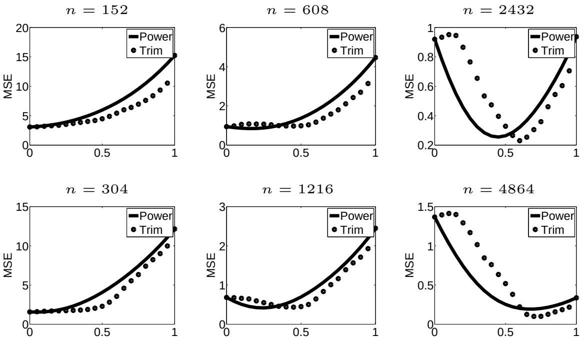

Figure 3: Age

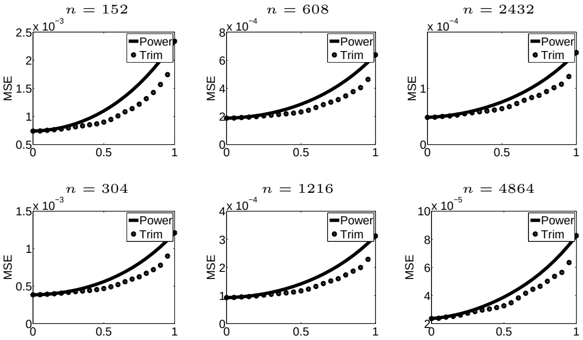

Figure 4: Major Depression

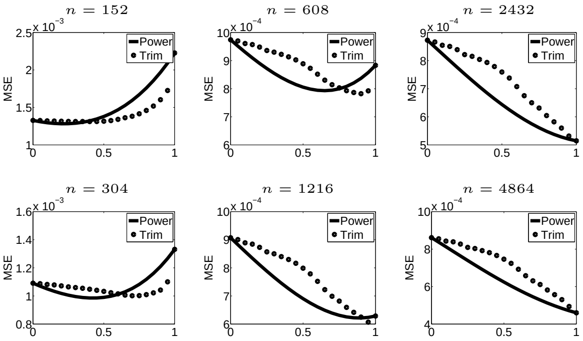

Figure 5: Substance Abuse

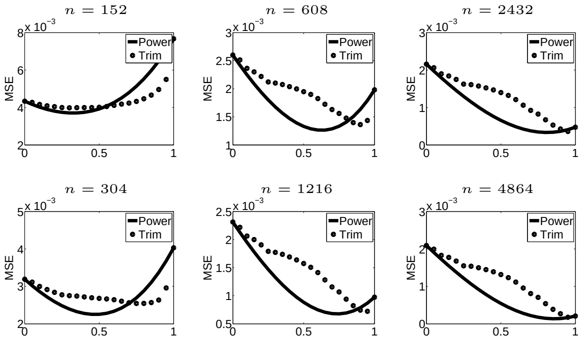

Figure 6: Gender

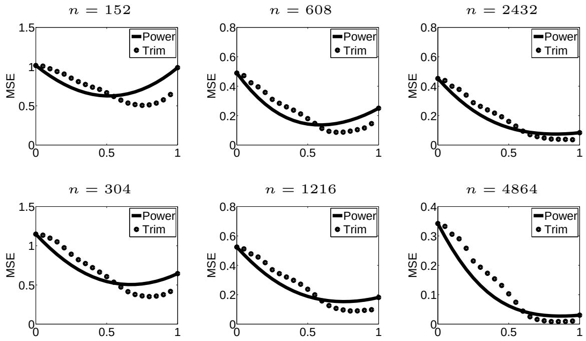

Figure 7: Education

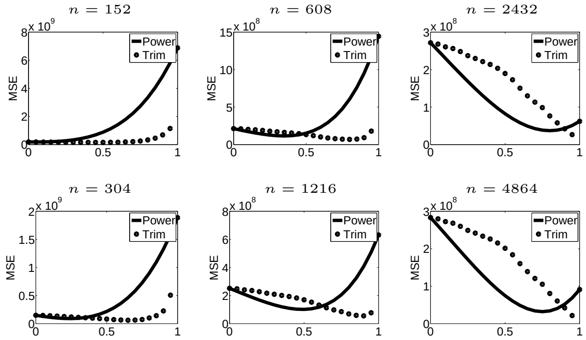

Figure 8: Survey Weight w

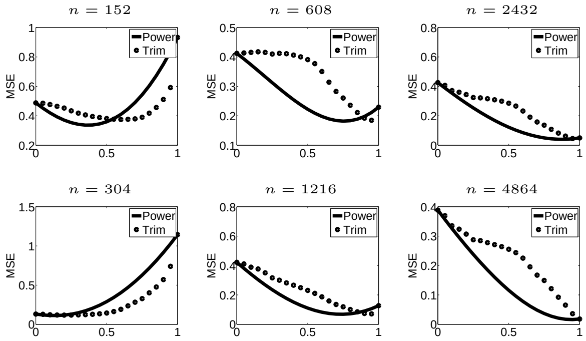

Figure 9: Height

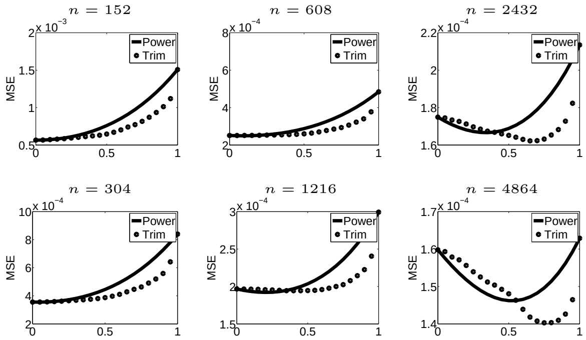

Figure 10: Social Phobia

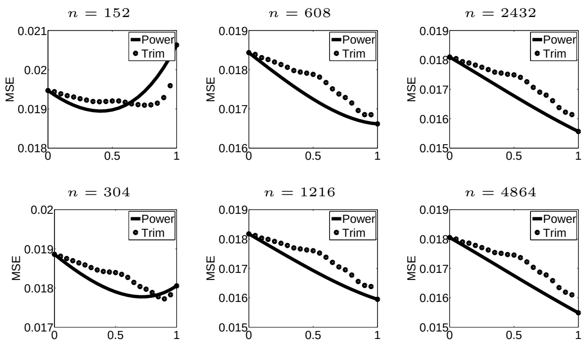

Figure 11: Immigrant

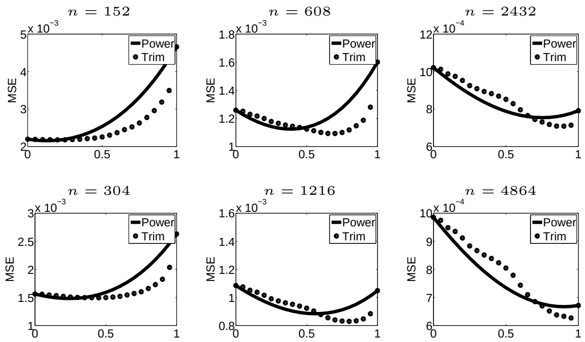

Figure 12: Any Disorder

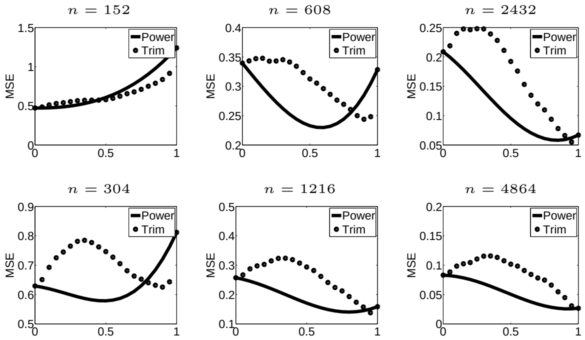

Figure 13: K10 Distress

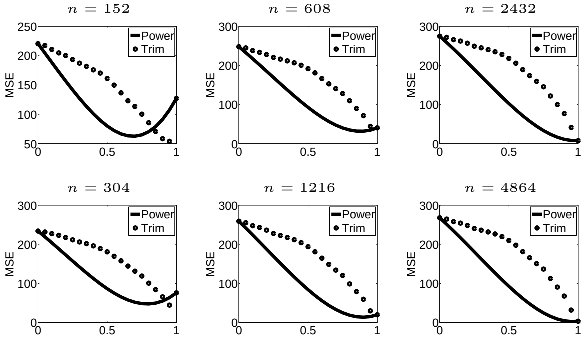

Figure 14: Agepluswgt