Blending of Probability and Convenience Samples:

Applications to a Survey of Military Caregivers

Michael Robbins
RAND Corporation

Collaborators:
Bonnie Ghosh-Dastidar, Rajeev Ramchand

September 25, 2017

Probability and Convenience Samples

It may be of interest to draw inferences regarding a rare segment of a population.

- E.g., what are the needs of servicemembers of post-9/11 US military operations?

Probability sampling (the gold standard) is often infeasible for studying such segments

- Rare segments usually have no complete sampling frame.
- Existing representative panels will not have sufficient individuals from the segment.

Convenience sampling is an efficient, cost-effective alternative.

- Easy to collect
- Not representative $\Rightarrow$ Lower quality data

Here, we explore the utility of convenience samples as a supplement to a (small) probability sample.

- Blend two samples: Probability sample + Convenience sample

Blending via weighting

As is the common practice in survey analysis, representativeness of non-representative data is obtained through weighting

- A sampled individual i is given a weight w_i.

Weights are set as equal to the inverse of the individual's probablity of being included in the sample; e.g., $w_i = P(i \in S|\mathbf{x}_i)^{-1}$

- Conditional on an observed set of auxiliary variables ($\mathbf{x}_i$)

Do we calculate

- Sampling probabilities for the probability sample, i.e., $P(i \in S_1|\mathbf{x}_i)$, and the convenience sample, i.e., $P(i \in S_2|\mathbf{x}_i)$,
- Or sampling probabilities for the blended sample, i.e., $P(i \in S|\mathbf{x}_i)$, where $S = S_1 \cup S_2$?

Can do both:

- The former – *Disjoint blending* (both samples are representative individually)
- The latter – *Simultaneous blending* (the samples are representative only when combined).

Propensity Scores

For individual i,

- We do not know the probability of inclusion into the convenience sample (S_2) or the blended sample (S).
- We know the prob. of incl. into the probability sample (S_1).
  - This is written: $d_i := P(i \in S_1 | \mathbf{x}_i)$
- We can calculate the probability of inclusion into S_2 given that the individual has been included in one of the samples.

$$\gamma_i := P(i \in S_2 | i \in S_1 \cup S_2, \mathbf{x}_i)$$

  - These are the propensity scores
  - Note that γ_i can be estimated using a logistic model.

Some math shows that

$$P(i \in S_2 | \mathbf{x}_i) = \frac{d_i \gamma_i}{1 - \gamma_i}$$

and

$$P(i \in S | \mathbf{x}_i) = \frac{d_i}{1 - \gamma_i}$$

Inverse probability weighting

Simultaneous weights

- Set $w_i = P(i \in S|\mathbf{x}_i)^{-1} = {d_i}^{-1}(1 - \gamma_i)$ for $i \in S$.

Note that this requires d_i for $i \in S_2$. How to calculate?

For probability samples, sampling probabilities are selected by the analyst for $i \in \Omega$.

- E.g., 100 individuals from a stratum of size 1000 are sampled
- Each individual in the stratum has a 10% probability of being sampled (even those not sampled).

Even the probability sample is likely subject to non-response

- Denote those sampled for S_1 by S_1^* and set $d_i^* = P(i \in S_1^*, \mathbf{x}_i)$.
- Let $r_i = P(i \in S_1 | i \in S_1^*, \mathbf{x}_i) = 1/(1 + \exp\{-\alpha_0 - \boldsymbol{\alpha}'\mathbf{x}_i\})$.
- Therefore, $d_i = d_i^* r_i$ — can calculate for all $i \in \Omega$.

Inverse probability weighting

Disjoint weights:

- Each sample uses its respective inclusion probabilities.

Consider:

- If we set $w_i = P(i \in S_1|\mathbf{x}_i)^{-1}$ for $i \in S_1$, then S_1 is representative of the population.
- Similarly, if we set $w_i = P(i \in S_2|\mathbf{x}_i)^{-1}$ for $i \in S_2$, then S_2 is representative of the population.

A separate set of blending weights can be calculated by

- setting $w_i = \kappa P(i \in S_1|\mathbf{x}_i)^{-1}$ for $i \in S_1$
- and $w_i = (1-\kappa)P(i \in S_2|\mathbf{x}_i)^{-1}$ for $i \in S_2$.
- κ: how much emphasis to give one sample over the other
  - We suggest picking $\kappa \in [0, 1]$ to minimize the design effect.

Calibration

Assume that population totals for the set of auxiliary variables $\mathbf{x}_i$,

$$\mathbf{t}_x = \sum_{i \in \Omega} \mathbf{x}_i,$$

are known. In this case, calibration is applicable.

- If $\mathbf{t}_x$ is unknown, it can be approximated using the probability sample: $\hat{\mathbf{t}}_x = \sum_{i \in S_1} d_i^{-1} \mathbf{x}_i$

Simultaneous calibration: Solve for weights $\{w_i\}$ that satisfy

$$\mathbf{t}_x = \sum_{i \in S} w_i \mathbf{x}_i$$

Disjoint calibration: Solve for weights $\{w_{1,i}\}$ & $\{w_{2,i}\}$ that satisfy

$$\mathbf{t}_x = \sum_{i \in S_1} w_{1,i} \mathbf{x}_i \quad \text{and} \quad \mathbf{t}_x = \sum_{i \in S_2} w_{2,i} \mathbf{x}_i,$$

and set $w_i = \kappa w_{1,i}$ for $i \in S_1$ and $w_i = (1 - \kappa) w_{2,i}$ for $i \in S_2$. The optimal choice of κ is as before.

Assumptions

As in propensity scores, our methods require an ignorability (or exchangability assumption).

- No unknown confounders
- That is, the set of auxiliary variables is sufficient for modeling differences between the two samples.

Could circumvent this by using outcome as an auxiliary variable

- You don't want to do this... will explain later

Assumptions

Can test this assumption (i.e., is blending adequate)

- Compare weighted outcome between the two samples
- Should be similar
- Must use disjoint blending

We test the adequacy of blending (with respect to y_i) by fitting the model

$$y_i = \mu + \delta 1_{\{i \in S_2\}} + \epsilon_i,$$

for $i \in S$ using weighted least squares (with the disjoint weights). $1_{\{A\}}$ represent the indicator of event A. We test

$$\mathcal{H}_0 : \delta = 0 \quad \text{against} \quad \mathcal{H}_1 : \delta \neq 0.$$

Military Caregivers

Our methods are applied to a survey of military caregivers

- Population includes unpaid caregivers of wounded servicemembers.
- Here, we focus on caregivers of post-9/11 servicemembers

The military caregivers project (Ramchand et al., 2014)

- Sponsored by the Elizabeth Dole Foundation
- Assess the needs of and difficulties encountered by military caregivers
- Compare pre- and post-9/11 military caregivers; compare all military caregivers to civilian caregivers, non-caregivers; etc.

Military Caregivers

The probability sample: KnowledgePanel (KP)

- The KnowledgePanel is a nationally representative panel of 45,000 American Adults (as of 8/2013).
- The whole panel was given a screener to determine if they were an unpaid caregiver
  - All military caregivers were sampled and surveyed
  - Veterans with unpaid caregivers were sampled and surveyed.
  - Some civilian caregiveres and non-caregivers were randomly sampled and surveyed.
  - This only yielded 72 post-9/11 military caregivers.

Stratum	KP
Veterans	251
Post-9/11 Caregiver	72
Pre-9/11 Caregiver	522
Civilian Caregiver	1828
Non-Caregiver	1163

Table: Respondents by stratum

Military Caregivers

The convenience sample: Wounded Warrior Project (WWP)

- WWP: A non-profit veteran's service that offers support to servicemembers injured after 9/11/2001.
- The WWP maintains a database of caregivers of veterans.
  - The caregivers contained therein are not thought to be representative of all post-9/11 caregivers.
- This database was used as a sampling frame for our convenience sample

Stratum	KP	WWP
Veterans	251	–
Post-9/11 Caregiver	72	281
Pre-9/11 Caregiver	522	3
Civilian Caregiver	1828	–
Non-Caregiver	1163	–

Table: Respondents by stratum

Military Caregivers

We calculate three sets of weights:

- Simultaneous propensity scores (SPS)
  - Propensity scores were calculated using a logistic model
- Disjoint propensity scores (DPS)
- Simultaneous calibration (SC)
  - Benchmark values calculated using the the probability sample and reports from veterans when feasible.

Disjoint calibration weights are not calculated because there were not feasible weights to match the convenience sample to the benchmarks across the auxiliary variables.

Results: Auxiliary Variables

	KP Only		WWP Only		Blended: Unweighted	
Variable Description	Mean	s.e.	Mean	s.e.	Mean	s.e.
Caregiver lives w/ veteran	0.453	0.089	0.858	0.021	0.755	0.032
Vet deployed to war zone†	0.577	0.090	0.929	0.015	0.839	0.029
Vet: rating 70+%	0.280	0.080	0.715	0.027	0.605	0.033
Vet: traumatic brain injury	0.164	0.073	0.687	0.028	0.554	0.032
Vet: m.h. problems†	0.450	0.088	0.897	0.018	0.783	0.033
Cg: is female†	0.476	0.089	0.940	0.014	0.822	0.032

Table: Some of the auxiliary variables used for blending

Results: Auxiliary Variables

Variable Description	KP Only		WWP Only		Blended: Unweighted	
	Mean	s.e.	Mean	s.e.	Mean	s.e.
Caregiver lives w/ veteran	0.453	0.089	0.858	0.021	0.755	0.032
Vet deployed to war zone†	0.577	0.090	0.929	0.015	0.839	0.029
Vet: rating 70+%	0.280	0.080	0.715	0.027	0.605	0.033
Vet: traumatic brain injury	0.164	0.073	0.687	0.028	0.554	0.032
Vet: m.h. problems†	0.450	0.088	0.897	0.018	0.783	0.033
Cg: is female†	0.476	0.089	0.940	0.014	0.822	0.032

Table: Some of the auxiliary variables used for blending

Results: Auxiliary Variables

Variable Description	KP Only		WWP Only		Blended: Unweighted	
	Mean	s.e.	Mean	s.e.	Mean	s.e.
Caregiver lives w/ veteran	0.453	0.089	0.858	0.021	0.755	0.032
Vet deployed to war zone†	0.577	0.090	0.929	0.015	0.839	0.029
Vet: rating 70+%	0.280	0.080	0.715	0.027	0.605	0.033
Vet: traumatic brain injury	0.164	0.073	0.687	0.028	0.554	0.032
Vet: m.h. problems†	0.450	0.088	0.897	0.018	0.783	0.033
Cg: is female†	0.476	0.089	0.940	0.014	0.822	0.032

Table: Some of the auxiliary variables used for blending

Results: Auxiliary Variables

	Bench-marks	Blended: Weighted		
		SPS	DPS	SC
Caregiver lives w/ veteran	0.453	0.467	0.573	0.489
Vet deployed to war zone†	0.559	0.592	0.631	0.591
Vet: rating 70+%	0.280	0.275	0.320	0.308
Vet: traumatic brain injury	0.164	0.133	0.166	0.189
Vet: m.h. problems†	0.510	0.491	0.529	0.549
Cg: is female†	0.555	0.509	0.557	0.592

Table: Some of the auxiliary variables used for blending

- SPS – Simultaneous propensity scores
- DPS – Disjoint propensity scores
- SC – Simultaneous calibration

Military Caregivers

Comments on the previous table:

- The probability and convenience sample (unweighted) give very different results.
  - The WWP has a much higher portion of females and caregivers of severely disabled veterans.
- The unweighted blended data have a substantial decrease in precision over the KP alone, but will clearly be biased
- The weighted blended data satisfy the benchmarks (though less so for the DPS weights)
- Simultaneous blending yields smaller standard errors than disjoint blending.

Results: Outcomes Variables

Variable Description	KP Only	WWP Only	Blending Weights SPS	DPS	SC
Caregiver depression	7.071	9.485	7.661	7.574	7.801
Caregiver anxiety	38.41	50.90	41.55	43.66	41.71
Quit work entirely	0.204	0.502	0.206	0.237	0.237
Caregiving disturbs sleep	0.474	0.808	0.499	0.536	0.554
Caregiver financial strain	0.588	0.762	0.593	0.627	0.575
Caregiver feels overwhelmed	0.398	0.815	0.443	0.480	0.480

Table: Results of blending for outcome variables: **Means**

- SPS – Simultaneous propensity scores
- DPS – Disjoint propensity scores
- SC – Simultaneous calibration

Results: Outcomes Variables

Variable Description	KP Only	WWP Only	Blending Weights SPS	DPS	SC
Caregiver depression	0.751	0.393	0.590	0.743	0.574
Caregiver anxiety	5.104	1.719	3.849	4.405	3.646
Quit work entirely	0.071	0.030	0.041	0.054	0.047
Caregiving disturbs sleep	0.091	0.024	0.067	0.069	0.062
Caregiver financial strain	0.085	0.025	0.064	0.064	0.060
Caregiver feels overwhelmed	0.087	0.023	0.064	0.068	0.061

Table: Results of blending for outcome variables: **S.E.'s**

- SPS – Simultaneous propensity scores
- DPS – Disjoint propensity scores
- SC – Simultaneous calibration

Results: Outcomes Variables

Variable Description	KP Mean	WWP Mean	Blending Weights SPS	DPS	SC
Caregiver depression	7.071	9.485	0.043	0.399	0.052
Caregiver anxiety	38.41	50.90	0.036	0.107	0.260
Quit work entirely	0.204	0.502	0.004	0.369	0.046
Caregiving disturbs sleep	0.474	0.808	0.002	0.210	0.019
Caregiver financial strain	0.588	0.762	0.088	0.372	0.163
Caregiver feels overwhelmed	0.398	0.815	0.000	0.114	0.000

Table: Results of blending for outcome variables: **p-values**

- SPS – Simultaneous propensity scores
- DPS – Disjoint propensity scores
- SC – Simultaneous calibration

Military Caregivers

Comments on the previous table:

- The probability and convenience sample (unweighted) give very different results for outcome variables as well.
  - Caregivers in the WWP endure more hardships.
  - This is mainly evident in depression and anxiety levels
- Using weighted blended data, one estimates that caregivers endure higher levels of hardships than one would estimate if one used only the KP.
- The weighted blended have higher precision than the KP only, though not to the same degree as the unweighted blended data
- Blending is adequate for most outcomes

A *post hoc* blending estimator

Recall that disjoint weights can be used to find

- $\hat{\theta}_1$ – An unbiased estimator of θ found with $i \in S_1$
- $\hat{\theta}_2$ – An unbiased estimator of θ found with $i \in S_2$

Many authors suggest *post hoc* blending estimation:

$$\bar{\theta} = \bar{\kappa}\hat{\theta}_1 + (1 - \bar{\kappa})\hat{\theta}_2$$

$\bar{\kappa}$ is selected to minimize the mean squared error of $\bar{\theta}$

$$\bar{\kappa} = \mathrm{Var}(\hat{\theta}_2)/(\mathrm{Var}(\hat{\theta}_1) + \mathrm{Var}(\hat{\theta}_2))$$

Involves recalculation of $\bar{\kappa}$ for each estimator.

Simulations

Use the real caregiver data to develop a pseudo-population of caregivers

- 940 pseudo-post-9/11 caregivers
- 1806 pseudo-pre-9/11 caregivers

Draw samples (to be blended) from the pseudo-population:

- Draw a probability sample of 75 pseudo-post-9/11 caregivers at random.
- Draw a convenience sample of pseudo-post-9/11 caregivers according to a logistic model:

$$\log\left(\frac{\rho_i}{1-\rho_i}\right) = b_0 + \boldsymbol{b}_1'\mathbf{v}_i, \tag{1}$$

Simulations

Consider four settings for choices of ρ_i

- Setting 1: ρ_i depends only on the auxiliary variables
- Setting 2: ρ_i depends on auxiliary variables and latent (non-outcome) variable
- Setting 3: ρ_i depends on auxiliary variables and the outcome variable
- Setting 4: ρ_i depends on auxiliary variables and the outcome variable, while the set of auxiliary variables has been expanded to include an additional variable related to the outcome.

Simulations

Variable	Value of coefficient in (1)			
	Setting 1	Setting 2	Setting 3	Setting 4
Intercept	$-\log(2)$	$-\log(2)$	$-\log(2)$	$-\log(2)$
Caregiver depression	0^*	0^*	τ^*	τ^*
Caregiver anxiety	0^*	τ^*	0^*	0
Caregiver gender	4/3	4/3	4/3	4/3
Caregiver lives w/ care recipient	1/3	1/3	1/3	1/3
Care recipient is single	1/3	1/3	1/3	1/3
Vet. deployed to a war zone	1/3	1/3	1/3	1/3
Vet. has disability rating	1	1	1	1
Vet. has disability rating of 70+%	1	1	1	1
Vet. has service-related TBI	1	1	1	1

Table: Coefficient values for the mechanism used to draw the convenience sample—τ is a positive constant that is used as a tuning parameter in order to vary the degree to which the samples are differentiated. An asterisk indicates the variable is *not* used as an auxiliary variable in the calculation of weights within the respective setting.

Simulations

The process is replicated K times. Let

- $\hat{\theta}$ – benchmark value of a parameter θ

Quantities tracked for each replication include:

- $\hat{\theta}^{[k]}$ – estimated value of θ in replication k
- $p^{[k]}$: the p-value of a test of the adequacy of blending
- $\text{DEFF}^{[k]}$: The design effect when estimating $\hat{\theta}^{[k]}$

Quantities reported (aggregated across replications):

- Bias $= \frac{1}{K}\sum_{k=1}^{K}[100(\hat{\theta}^{[k]} - \hat{\theta})/\hat{\theta}]$
- root-MSE $= \sqrt{\frac{1}{K}\sum_{k=1}^{K}[100(\hat{\theta}^{[k]} - \hat{\theta})/\hat{\theta}]^2}$
- Rejection rate: $\hat{p} = \frac{1}{K}\sum_{k=1}^{K} 1_{\{p^{[k]} \leq \alpha\}}$
- Design effect $= \frac{1}{K}\sum_{k=1}^{K} \text{DEFF}^{[k]}$

Simulations

Methods used include:

- The probability sample only (KP)
- Unweighted blended samples (unw)
- Simultaneous propensity scores (SPS)
- Disjoint propensity scores (DPS)
- *Post hoc* blending w/ propensity scores ($\bar{\kappa}PS$)
- Simultaneous calibration (SC)
- Disjoint calibration (DC).
- *Post hoc* blending w/ calibration ($\bar{\kappa}C$)

Simulations

		KP	unw	SPS	DPS	$\bar{\kappa}$PS
Setting 1	DEFF	1.00	1.00	1.66	1.85	—
	Bias	-0.10	12.61	0.06	1.73	1.51
	rMSE	9.98	12.87	5.86	6.45	6.38
	Rejection rate	—	0.294	0.259	0.049	—
Setting 2	DEFF	1.00	1.00	1.68	1.86	—
	Bias	-0.12	19.17	7.84	10.10	9.82
	rMSE	9.95	19.34	9.79	11.93	11.74
	Rejection rate	—	0.544	0.606	0.241	—
Setting 3	DEFF	1.00	1.00	1.70	1.90	—
	Bias	0.06	23.39	13.19	15.93	15.41
	rMSE	9.97	23.53	14.49	17.25	16.85
	Rejection rate	—	0.696	0.795	0.488	—
Setting 4	DEFF	1.00	1.00	1.69	1.91	—
	Bias	-0.07	23.36	7.94	10.96	10.65
	rMSE	9.75	23.50	11.01	13.63	13.49
	Reject rate	—	0.706	0.772	0.232	—

Table: Results for mean depression levels in the pseudo-post-9/11 caregivers. $K = 10,000$ replications are used.

Simulations

		KP	unw	SC	DC	$\bar{\kappa}$C
Setting 1	DEFF	1.00	1.00	1.75	2.26	—
	Bias	-0.10	12.61	-0.26	0.20	-0.16
	rMSE	9.98	12.87	6.04	6.64	6.65
	Rejection rate	—	0.294	0.304	0.041	—
Setting 2	DEFF	1.00	1.00	1.77	2.26	—
	Bias	-0.12	19.17	7.17	8.01	7.81
	rMSE	9.95	19.34	9.40	10.59	10.49
	Rejection rate	—	0.544	0.634	0.178	—
Setting 3	DEFF	1.00	1.00	1.79	2.30	—
	Bias	0.06	23.39	12.28	13.70	13.28
	rMSE	9.97	23.53	13.77	15.50	15.25
	Rejection rate	—	0.696	0.800	0.422	—
Setting 4	DEFF	1.00	1.00	1.78	2.35	—
	Bias	-0.07	23.36	7.45	8.05	7.76
	rMSE	9.75	23.50	10.75	11.63	11.54
	Reject rate	—	0.706	0.802	0.124	—

Table: Results for mean depression levels in the pseudo-post-9/11 caregivers. $K = 10,000$ replications are used.

Simulations

Conclusions from the previous table:

- Weighted blending is always preferable to unweighted blending.
- When blending is adequate (Setting 1)
  - Weighted blending is preferable to using only the convenience sample.
  - p-values of the test for adequacy of blending are close to their nominal levels.
- Calibration and propensity scores yield similar results
  - PS-based methods yield lower design effects
  - Calibration observe lower rMSE
- Simultaneous blending yields lower design effects than disjoint blending.
- Only disjoint blending can be used to assess adequacy.

Simulations

Conclusions from the previous table:

- Even though anxiety is not the outcome of interest, we see in Setting 2 that allowing the probability of selection into the convenience sample to depend upon anxiety induces bias into estimators found by the weighting schemes. However, this bias would be smaller if depression and anxiety were not highly correlated.
- Similarly, Setting 4 (when compared to Setting 3) illustrates that if the probability of selection into the convenience sample depends on the outcome of interest (depression), bias can be reduced by using additional variables that are correlated with the outcome as auxiliary variables in the calculation of weights.

Variance Estimation

We've skirted the issue of variance estimation to this point.

Can traditional methods for variance estimation in survey data be used with the blending weights?

- Linearization – Complex algebraic formulas for estimation variance with weighted data
- Jackknife – Data are segmented into replicate groups.
  - Weights are recalculated for each replicate group (i.e., the weighting algorithm is re-run).

Variance Estimation: *Simulation*

Generate synthetic probability and convenience samples

One outcome

- Goal is to estimate the mean of this outcome with blended data

Two auxiliary variables

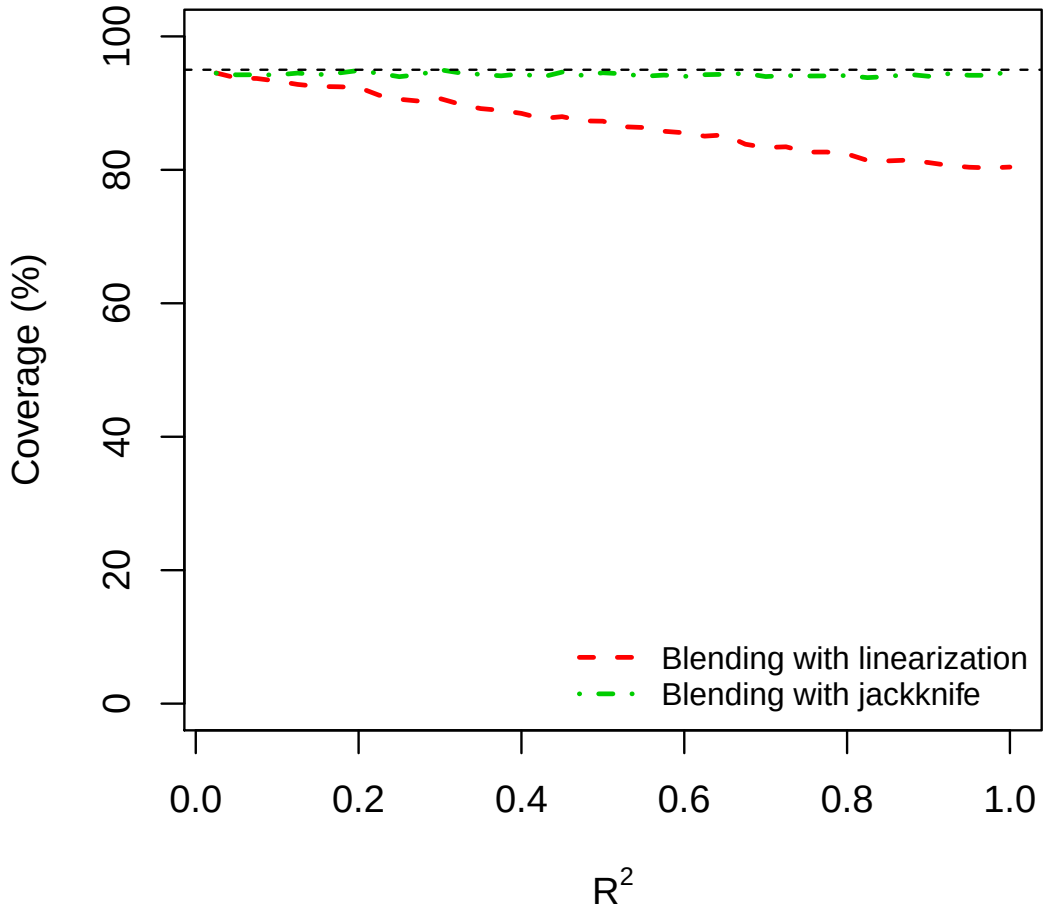
- Probability of selection into the convenience sample depends upon these variables

We vary the degree to which the auxiliary variables predict the outcome variable (R^2).

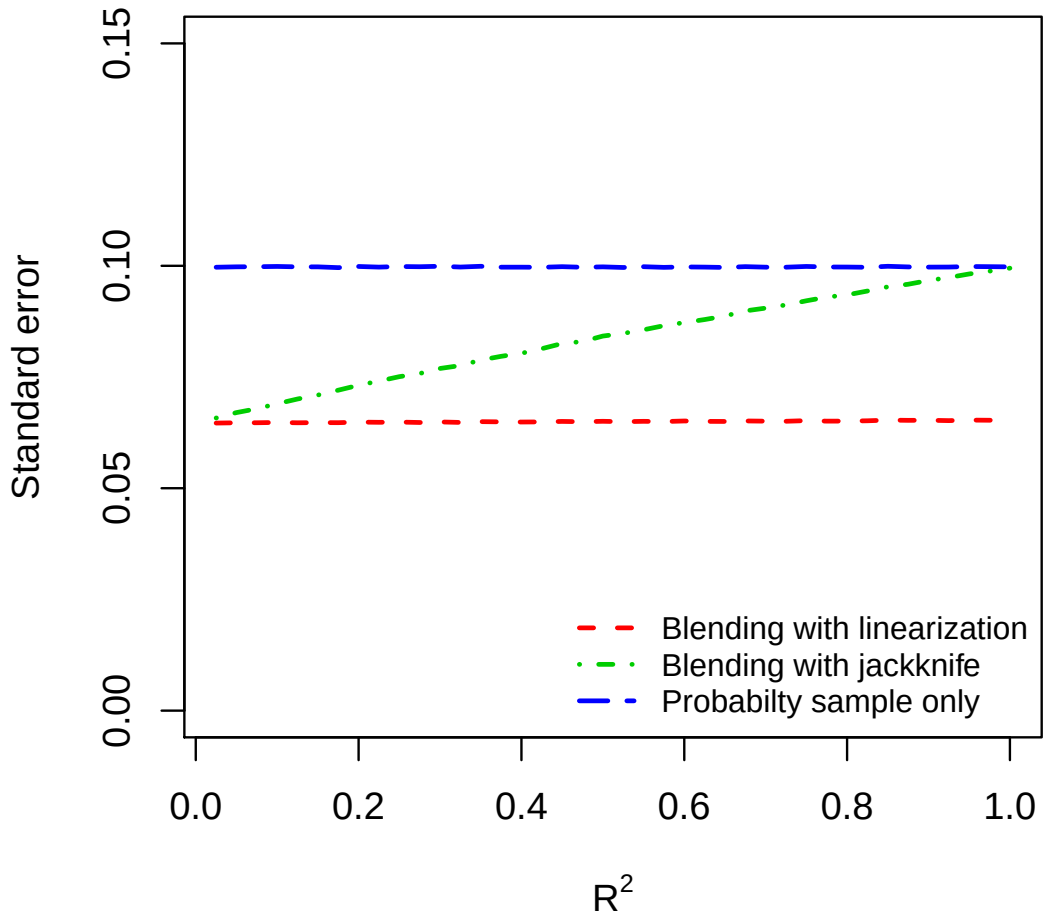
For different methods of variance estimation, we calculate

- Estimated standard error of the outcome's mean
- Coverage – portion of simulations in which the true mean falls in the estimated 95% confidence interval

Variance Estimation: Coverage



Variance Estimation: Standard Error



The more predictive the auxiliary variables are of the outcome, the less efficiency is gained by using the convenience sample

Summary

We introduced four methods for calculating blending weights.

- An important discrepancy is made between simultaneous weights and disjoint weights
  - Simultaneous weights give smaller design effects
  - Disjoint weights are used to assess the adequacy of blending.

Is blending with convenience samples too dangerous?

- The assumptions and limitations are the same as propensity score methods
- People do that all the time so why not do this?

Key question: Should you be exhaustive in your selection of auxiliary variables?

- Advantages: Reduces the potential for sampling bias (due to unknown confounders)
- Disadvantages: Reduces the gain in efficiency from the use of the convenience sample.