Using bootstrapping methods to determine difference in effects

Question 1: When using a uniform set of predictors, are the effects larger when using the full population than the effect when using a sibling subsample?

Background: As stated by Jaffee et al. (2012), observational studies using fixed effect methods to compare siblings within families are considered to be a stronger research design than population studies by eliminating "genetic confounds due to passive gene-environment correlation (p.275)" and controlling for reciprocal causation (Lahey & D'Onofio, 2010). When examining the causal effects of risk factors, stronger research designs tend to show smaller effects (Jaffee et al, 2012).

Example: What Predictors Matter: Risk Factors for Cognitive and Emotional Development paper (to be submitted shortly) by Wall-Wieler, Roos, Chateau and Rosella.

Four late adolescent outcomes are being measured – High School Graduation, ADHD, Conduct Disorder/ODD and Juvenile Bipolar. The full population has 62 739 individuals. A subsample of two randomly selected consecutive siblings was selected to run an additional analysis and to determine whether the results from the larger sample (population) were robust. There are slight differences in the model. Both models include indicator variables on birth order; the siblings sample does not include the indicator variable 'only child'. The sibling models also include four variables examining sibling spacing, which are excluded in the population models. Aside from these small differences, the models are the same; they share 31 variables. Since the sibling models are considered to be a stronger research design since they are able to eliminate potential familial confounding, the effects are expected to be smaller. To determine whether this is the case here, we tested to see whether there is a significant difference between the effects seen in the population model and the effects seen in the sibling model.

Method

1) Get effects for the population model and the sibling model.
2) Calculate the difference between the common effects (diff_estimate = estimate_pop – estimate_sib)
3) Use unrestricted random sampling with replacement 500 times with different randomly selected samples to get the bootstrapped estimate of the standard error of the difference. (std_err_diff)
4) Compute a 95% confidence interval around the difference (diff_estimate – 1.96*std_err_diff, diff_estimate + 1.96*std_err_diff)
5) If the confidence interval includes 0, then there is no significant difference.

SAS Code

```
/*read in population and siblings files, make sure there are no duplicates*/
proc sort data = project.population_filename out = population nodupkey; by scrphin; run;
proc sort data = project.siblings_filename out = siblings nodupkey; by scrphin; run;

/*get estimates of the effects for population sample*/
	ods output glimmix.parameterestimates= pop_estimate;
	proc glimmix data = population noclprint noitprint ic=q;
		class sib_group;
		model grad_4yr = std_sefi rural mage male bwt_lt2500 bwt_2500_3500
		orderfam1 orderfam3 orderfam4 orderfam5 orderfam6 orderfam7
		orderfam8 orderfam9 orderfam10 fsm0013_div fsm0013_death
		fsm0013_twor  fsu0013_zero fsu0013_mar fsu0013_two
		moves adhd_0_13 conduct_odd_0_13 bipolar_0_13 asthma_0_13 maj_inj_0_13
		maj_cond_0_13 childcare0913 received0913 sa_use_9_13 g9
		/ dist=binary link=logit ddfm=bw solution;
		Random intercept / subject = sib_group;
	run;
	ods output close;

/*get estimates of the effects for sibling sample*/
	ods output glimmix.parameterestimates = sib_estimate;
	proc glimmix data = siblings noclprint noitprint ic=q;
		class sib_group;
		model grad_4yr = std_sefi rural mage diffbd_q1 diffbd_q2 diffbd_q3 diffbd_q4
```

```
                male bwt_lt2500 bwt_2500_3500 orderfam3 orderfam4 orderfam5 orderfam6
                orderfam7 orderfam8 orderfam9 orderfam10 fsm0013_div fsm0013_death
                fsm0013_twor fsu0013_zero fsu0013_mar fsu0013_two
                moves adhd_0_13 conduct_odd_0_13 bipolar_0_13 asthma_0_13 maj_inj_0_13
                maj_cond_0_13 childcare0913 received0913 sa_use_9_13 g9
                / dist=binary link=logit ddfm=bw solution;
                Random intercept / subject = sib_group;
        run;
        ods output close;

/*get difference in estimate*/
data pop_estimate (keep = effect pop_estimate);
        Set pop_estimate;
        Pop_estimate = estimate;
run;

proc sort data = pop_estimate; by effect; run;

data sib_estimate (keep = effect sib_estimate);
        set sib_estimate;
        sib_estimate = estimate;
run;

proc sort data = sib_estimate; by effect; run;

data estimate_diff;
        merge pop_estimate (in = a) sib_estimate (in = b);
        by effect;
        if a and b;
run;

data estimate_diff;
        set estimate_diff;
        diff_estimate = pop_estimate – sib_estimate;
run;

/*when we are randomly sampling from the siblings sample, we need to make sure that we
randomly selecting siblings, not individuals  - create two datasets, each dataset has one sibling
from each family, if we randomly select siblings based on the same seed, our randomly selected
sample will include the same sib_groups*/

proc sort data = siblings;
        by sib_group;
run;

data family_first;
```

```
	set siblings;
	if first.sib_group;
run;

data family_last;
	set siblings;
	if last.sib_group;
run;

/*****************************************************************
HS Grad Estimate Bootstrap – Difference in Estimate
*****************************************************************/

ods listing close;
%macro boostrap;
	options nonotes;
	data grad_pop_bstrap; run;   /*blank population dataset*/
	data grad_sib_bstrap; run;    /*blank sibling dataset*/
	data grad_diff_bstrap; run;   /*blank differences dataset*/

	%do e = 1 %to 500;

	/*population model*/
	proc surveyselect data = population out = bootstrap_population seed = 1789011&e
		method = urs sampsize=62739 outhits rep=1 noprint;
	run;

	ods output glimmix.parameterestimates=bout_pop;
	proc glimmix data = bootstrap_population noclprint noitprint ic=q;
		class sib_group;
		model grad_4yr = std_sefi rural mage male bwt_lt2500 bwt_2500_3500
		orderfam1 orderfam3 orderfam4 orderfam5 orderfam6 orderfam7
		orderfam8 orderfam9 orderfam10 fsm0013_div fsm0013_death
		fsm0013_twor  fsu0013_zero fsu0013_mar fsu0013_two
		moves adhd_0_13 conduct_odd_0_13 bipolar_0_13 asthma_0_13 maj_inj_0_13
		maj_cond_0_13 childcare0913 received0913 sa_use_9_13 g9
		/ dist=binary link=logit ddfm=bw solution;
		Random intercept / subject = sib_group;
	run;
	ods output close;

	data grad_pop_bstrap (keep = effect pop_estimate);
		set grad_pop_bstrap bout_pop (keep = estimate effect);
		pop_estimate = estimate;
	run;
```

```
/*sibling model*/
proc surveyselect data = family_first out=bootstrap_1 seed = 1789011&e
        method = urs samprate = 1 outhits rep=1 noprint;
run;

proc surveyselct data = family_last out=bootstrap_2 seed = 1789011&e
        method = urs samprate = 1 outhits rep=1 noprint;
run;

data bootstrap_siblings;
        set bootstrap_1 bootstrap_2;
run;


ods output glimmix.parameterestimates = bout_sib;
proc glimmix data = bootstrap_siblings noclprint noitprint ic=q;
        class sib_group;
        model grad_4yr = std_sefi rural mage diffbd_q1 diffbd_q2 diffbd_q3 diffbd_q4
        male bwt_lt2500 bwt_2500_3500 orderfam3 orderfam4 orderfam5 orderfam6
        orderfam7 orderfam8 orderfam9 orderfam10 fsm0013_div fsm0013_death
        fsm0013_twor fsu0013_zero fsu0013_mar fsu0013_two
        moves adhd_0_13 conduct_odd_0_13 bipolar_0_13 asthma_0_13 maj_inj_0_13
        maj_cond_0_13 childcare0913 received0913 sa_use_9_13 g9
        / dist=binary link=logit ddfm=bw solution;
        Random intercept / subject = sib_group;
run;
ods output close;

data grad_sib_bstrap (keep = effect sib_estimate);
        set grad_sib_bstrap bout_sib (keep = effect estimate);
        sib_estimate = estimate;
run;

proc sort data = grad_pop_bstrap; by effect; run;
proc sort data = grad_sib_bstrap; by effect; run;

data grad_diff;
        merge grad_pop_bstrap (in = a) grad_sib_bstrap (in = b);
        by effect;
        if a and b; /*only want to keep the common effects*/
run;

data grad_diff;
        set grad_diff;
        diff_estimate = pop_estimate – sib_estimate;
run;
```

```
	data grad_diff_bstrap (keep = effect diff_estimate);
		set grad_diff_bstrap grad_diff;
	run;

%end;
%mend bootstrap;
%bootstrap;

/*now we have 500 estimates of the difference between the effect in the population model and
the effect in the sibling model*/

data grad_diff_bstrap;
	set grad_diff_bstrap;
	where diff_estimate ^=.;
run;

proc sort data = grad_diff_bstrap; by effect; run;

proc transpose data = grad_diff_bstrap out = g1 prefix = diff_estimate;
	by effect;
run;

/*we can get the standard error of the difference by calculating the standard deviation of the 500
differences*/

data g1_means;
	set g1 (keep = effect diff_estimate:);
	n_stderr = std(of diff_estimate1 – diff_estimate500);
run;

/*get the estimate in difference from earlier calculation*/
Proc sort data = estimate_diff; by effect; run;

Proc sort data = g1_means; by effect; run;

Data confidence_interval;
	Merge estimate_diff (in = a) g1_means (in = b);
	By effect;
	If a and b;
Run;

Data confidence_interval;
	Set confidence_interval;
	LCL = diff_estimate – (1.96*n_stderr);
	UCL = diff_estimate + (1.96*n_stderr);
```

```
Run;

proc print data = confidence_interval;
        var effect LCL diff_estimate UCL;
        title ‘gradhs boostrapped std confidence interval of the difference’;
run;
```

Question 2: When using two distinct cohorts to model two distinct (but related) outcomes using a common set of predictors, are there differences in the effects?

Example: Predicting Three Types of Welfare—Where are the Differences? (to be submitted shortly) by Wall-Wieler, Roos, Chateau and Rosella.

The welfare program in Manitoba is the Employment and Income Assistance (EIA) program, and within this program, there are three main categories – General Assistance, Single Parent and Disability. Approximately 13 percent of individuals born between April 1, 1980 and November 30, 1987 received at least 2 months of EIA between 18 and 26. The objective of this paper was to determine whether there were differences in predictors for the three different types of EIA. Individuals who did not receive EIA in early adulthood were the 'controls' and were divided into three groups based on the proportion of EIA recipients using each type of EIA (Table 1).

Table 1 - Table 1 - Welfare use by Type

Welfare Type	Received Welfare, 18 - 25	Did not receive Welfare, 18 - 25	Total
General Assistance	1,913 (35.2%)	13,102	15 015
Single Parent	2,258 (41.5%)	15,485	17 743
Disability	1,217 (22.3%)	8,696	9 913

This gives us three distinct cohorts. Using a common set of predictors (mother's age at first birth, mother's marital status at birth, birth order, birth weight, sex, residential mobility 0 – 17, major mental health conditions 0 – 17, minor mental health conditions 0 – 17, major injuries 0 – 17), we were interested in determining whether there were significant differences between the effects. Some of the differences are easy to spot – a predictor will be significant for one welfare type but not another welfare type, or a predictor is a risk factor for one type and a protective factor for another. Other differences were not so easy to determine. If a predictor is a risk factor for two (or three) types of welfare, we needed to determine whether the predictor was a

significantly greater risk factor for one type of welfare. A very similar method as outlined above is used here. Only two types of welfare can be compared at one time. Will need to look at three separate differences (General Assistance – Disability; General Assistance – Single Parent; Disability – Single Parent). Below is the method to look at differences between General Assistance and Disability.

Method

1) Get effects for the general assistance model and the disability model.
2) Calculate the difference between the effects (diff_estimate = estimate_GA – estimate_D)
3) Use unrestricted random sampling with replacement 500 times with different randomly selected samples to get the bootstrapped estimate of the standard error of the difference. (std_err_diff) – See SAS code below
4) Compute a 95% confidence interval around the difference (diff_estimate – 1.96*std_err_diff, diff_estimate + 1.96*std_err_diff)
5) If the confidence interval includes 0, then there is no significant difference.

SAS Code

```
/*read in population and siblings files, make sure there are no duplicates*/
/*Get estimates from general assistance model*/
proc sort data = project.generalassistance_filename out = gen_assist nodupkey; by scrphin; run;

        ods output parameterestimates= ga_estimate;
        proc logistic data = gen_assist;
                model received_eia_18_25  =
                agefbirth_5 unmarried bw_1500 bw_1500_2500 bw_2500_3000 bw_3500
                b_order_2 b_order_3 b_order_4 b_order_5  male g9_index prev_moves
                prev_major_mental prev_minor_mental prev_major_injury;
        run;
        ods output close;

data ga_estimate (keep = ga_estimate variable);
        set ga_estimate;
```

```
	ga_estimate = estimate;
run;

proc sort data = project.disability_filename out = disability nodupkey; by scrphin; run;
/*Get estimates from general assistance model*/
	ods output parameterestimates= d_estimate;
	proc logistic data = disability;
		model received_eia_18_25  =
		agefbirth_5 unmarried bw_1500 bw_1500_2500 bw_2500_3000 bw_3500
		b_order_2 b_order_3 b_order_4 b_order_5  male g9_index prev_moves
		prev_major_mental prev_minor_mental prev_major_injury;
	run;
	ods output close;

data d_estimate (keep = d_estimate variable);
	set d_estimate;
	d_estimate = estimate;
run;

proc sort data = ga_estimate; by variable; run;
proc sort data = d_estimate; by variable; run;

data diff_estimate;
	merge ga_estimate d_estimate;
	by variable;
run;

data diff_estimate (keep = variable diff_estimate);
	set diff_estimate;
	diff_estimate = ga_estimate – d_estimate;
run;


/*****************************************************************
Standard Difference in Estimate
*****************************************************************/

ods listing close;
%macro boostrap;
	options nonotes;
	data gen_assist_bstrap; run;   /*blank general assistance dataset*/
	data disability_bstrap; run;    /*blank disability dataset*/
	data diff_bstrap; run;   /*blank differences dataset*/

	%do e = 1 %to 500;
```

```
/*general assistance model*/
proc surveyselect data = gen_assist out = bootstrap_gen_assist seed = 1789011&e
        method = urs sampsize=15015 outhits rep=1 noprint;
run;

ods output parameterestimates=bout_ga;
proc logistic data = bootstrap_gen_assist;
        model received_eia_18_25  =
        agefbirth_5 unmarried bw_1500 bw_1500_2500 bw_2500_3000 bw_3500
        b_order_2 b_order_3 b_order_4 b_order_5  male g9_index prev_moves
        prev_major_mental prev_minor_mental prev_major_injury;
run;
ods output close;

data gen_assist_bstrap (keep = variable ga_estimate);
        set gen_assist_bstrap bout_ga (keep = variable effect);
        ga_estimate = estimate;
run;

/*disability model*/
proc surveyselect data = disability out = bootstrap_disab seed = 1789011&e
        method = urs sampsize=9913 outhits rep=1 noprint;
run;

ods output parameterestimates=bout_dis;
proc logistic data = bootstrap_disab;
        model received_eia_18_25  =
        agefbirth_5 unmarried bw_1500 bw_1500_2500 bw_2500_3000 bw_3500
        b_order_2 b_order_3 b_order_4 b_order_5  male g9_index prev_moves
        prev_major_mental prev_minor_mental prev_major_injury;
run;
ods output close;

data disability_bstrap (keep = variable d_estimate);
        set disability_bstrap bout_dis (keep = variable estimate);
        d_estimate = estimate;
run;

proc sort data = gen_assist_bstrap; by variable; run;
proc sort data = disability_bstrap; by variable; run;

data eia_diff;
        merge gen_assist_bstrap (in = a) disability_bstrap (in = b);
        by variable;
        if a and b;
run;
```

```
	data eia_diff;
		set eia_diff;
		diff_estimate = ga_estimate – d_estimate;
	run;

	data diff_bstrap (keep = variable diff_estimate);
		set diff_bstrap eia_diff;
	run;

%end;
%mend bootstrap;
%bootstrap;

/*now we have 500 estimates of the difference between the effect in the general assistance model
and the effect in the disability model*/

data diff_bstrap;
	set diff_bstrap;
	where diff_estimate ^=.;
run;

proc sort data = diff_bstrap; by effect; run;

proc transpose data = diff_bstrap out = g1 prefix = diff_estimate;
	by effect;
run;

/*we can get the standard error of the difference by calculating the standard deviation of the 500
differences*/

data g1_means;
	set g1 (keep = effect diff_estimate:);
	n_stderr = std(of diff_estimate1 – diff_estimate500);
run;

/*get estimate of the difference calculated earlier*/
proc sort data = diff_estimate; by variable; run;
proc sort data = g1_means; by variable; run;

data confidence_interval;
	merge diff_estimate g1_means;
	by variable;
run;

data confidence_interval;
```

```
        set confidence_interval;
        LCL = diff_estimate – (1.96*n_stderr);
        UCL = diff_estimate + (1.96*n_stderr);
Run;

proc print data = confidence_interval;
        var variable LCL diff_estimate UCL;
        title ‘Boostrapped std err of the difference in estimates for general assistance and
disability’;
run;
```

References

Jaffee SR, Strait LB, Odgers CL. From correlates to causes: Can quasi-experimental studies and statistical innovations bring us closer to identifying the causes of antisocial behavior? *Psychol. Bull.* 2012:138(2):272-295.

Lahey BB, **D'Onofrio BM**. All in the family: Comparing siblings to test causal hypotheses regarding environmental influences on behavior. *Curr Dir Psychol.* 2010;19(5):319-323.