MWSUG 2017 – Paper AA08

Correcting for Selection Bias in a Clinical Trial

Shana Kelly, Spectrum Health: Healthier Communities, Grand Rapids, MI

ABSTRACT

Selection bias occurs when data does not represent the population intended, and balance across all potential confounding factors based on randomization did not happen. Selection bias can cause misleading results when doing statistical analysis, and should be corrected for. This paper explores a few alternative techniques to correct for a disparity between the various comparison groups in a clinical trial. Food Prescription is a small clinical trial conducted by Spectrum Health to get impoverished individuals in the Grand Rapids community with a chronic disease such as diabetes to consume more fresh fruits and vegetables. Health outcomes are compared between the treatment and control groups after taking into account all covariates. The procedures shown are produced using SAS® Enterprise Guide 7.1.

INTRODUCTION

This paper focuses on correcting for selection bias that occurred in a clinical trial, Food Prescription. Food Prescription was put on by Spectrum Health: Healthier Communities from July through October 2016. Participants were given a weekly voucher to use at a local farmers' market in Grand Rapids. All participants selected were also enrolled in Healthier Communities' Core Health program, which helps adults with a chronic disease such as diabetes or heart failure to self-manage their disease. The hypothesis with Food Prescription was that it would encourage low-income individuals to consume more fresh fruits and vegetables, while also supporting local agriculture. A pre-survey was given to get a baseline of how many servings of fruits and vegetables participants were eating before they started the Food Prescription program. A post-survey was given to assess how many servings of fresh produce individuals were now eating daily after twenty weeks of $20 food vouchers. Other outcome measures of interest were the number of days per week of at least thirty minutes of physical activity, a hemoglobin A1c level of less than 7.0% for diabetic participants, a Patient Activation Measure® (PAM-13) assessment score greater than 67, utilization of the emergency department, and being admitted as an inpatient. These measures were all taken from the data collected at the participants' home visits during the Core Health program, and hospital records. Other factors that were considered were race, age, gender, annual income, education level, and the participants' diagnosis of heart failure, diabetes, or both.

SELECTION BIAS

Selection bias occurs when data does not represent the population intended, and balance across all potential confounding factors based on randomization did not happen. Selection bias can cause misleading results when doing statistical analysis, and should be corrected for. The most severe bias was related to a large disparity in the race of individuals in the treatment and control groups. Figure 1 shows the proportion of each race in the treatment group, and Figure 2 shows the proportion of each race in the control group. The Hispanic population was oversampled in the treatment group, and the Caucasian population was oversampled in the control group.

Race of Participant-Treatment Group						
Race	White or Caucasian	African American	Hispanic	American Indian or Alaska Native	Unknown	Total
NUMBER	10	12	14	1	4	41
PERCENT	24.4	29.3	34.1	2.4	9.8	100

Figure 1: Race of participants in the treatment group

Race of Participant-Control Group							
Race	White or Caucasian	African American	Hispanic	Asian or Pacific Islander	American Indian or Alaska Native	Unknown	Total
NUMBER	8	5	5	1	1	1	21
PERCENT	38.1	23.8	23.8	4.8	4.8	4.8	100

Figure 2: Race of participants in the control group

This disproportionate breakdown of races between the two groups could mask the treatment effect, and should be corrected to obtain reliable results.

USING DEMOGRAPHIC FACTORS AS COVARIATES

The first method used to correct for selection bias was to create indicator variables for each race. The races with the highest frequency were white, African American, and Hispanic, so indicator variables were created for each of those. Separate models were built for each outcome variable of interest as the response. The difference in the number of servings of fresh fruits and vegetables the individual was eating daily from the beginning to the end of the twenty weeks was calculated, as well as the difference in the other outcome variables, and these differences were used in the models. Explanatory variables included all demographic variables, as well as the participants' diagnosis and which treatment or control group the individual was in. The three race indicator variables were used in the models instead of the overall race variable. The following code produces a backward-selected model to predict the difference in fruit and vegetable servings, and the output in Figure 3. This can be replicated for the other outcome variables.

```
proc glmselect data=final;
   class annual_income education_level Master_diag sex group white
   black(ref='0') hispanic(ref='0')/param=ref;
   model fruitdiff=Age_Enroll annual_income education_level Master_diag
   sex group white black hispanic/selection=backward(select=SL sls=.15)
   showpvalues;
run;
```

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-5.237473	2.432703	-2.15	0.0379
Age_Enroll	1	0.083454	0.041140	2.03	0.0497
Master_diag 1	1	0.383247	1.546263	0.25	0.8056
Master_diag 2	1	2.555198	1.106401	2.31	0.0266
hispanic 1	1	-1.347555	0.821573	-1.64	0.1094

Figure 3: Parameter estimates from model of difference in fruit and vegetable servings

Looking at the output above, it can be seen that the treatment versus control group is not a significant predictor in determining a participant's difference in fruit and vegetable consumption. Age and diagnosis have a positive relationship with the response, and the Hispanic indicator variable has a negative relationship. The coefficients can be interpreted as: for every one year increase in the enrollment age of a participant, their daily servings of fruits and vegetables increase by .08. Heart failure patients (Master_diag 1) have a predicted .38 daily increase in servings; whereas diabetic patients (Master_diag 2) have a 2.56 daily servings increase, compared to the reference category of patients diagnosed with

both diseases. Hispanic participants have a predicted decrease of 1.35 servings per day compared to those that are not Hispanic.

Figures 4 shows the parameter estimates of a linear model for the difference in 30-minute sessions of physical activity per week, produced by the code below.

```
proc glmselect data=final;
   class annual_income education_level Master_diag sex group white
   black(ref='0') hispanic(ref='0')/param=ref;
   model physdiff=Age_Enroll annual_income education_level Master_diag sex
   group white black hispanic/selection=backward(select=SL sls=.15)
   showpvalues;
run;
```

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.272628	1.811458	-1.25	0.2146
Age_Enroll	1	0.070017	0.035638	1.96	0.0542
white 1	1	-2.333736	0.717370	-3.25	0.0019

Figure 4: Parameter estimates from model of difference in 30 minute sessions of physical activity per week

Interpreting the parameter estimates above, for every one year increase in age of a participant, their estimated difference in number of thirty-minute physical activity sessions per week increases by .07. Race is a significant predictor in the model with white participants having an estimated decrease in physical activity by 2.3 sessions per week compared to non-white participants.

Measurements of participants' hemoglobin A1C were taken as close to enrollment and the end of the clinical trial as possible, and a difference was calculated and used as the linear model's response. For diabetes patients, an HbA1c level of less than 7% indicates that their diabetes is under control; therefore, a larger decrease is desirable. Figures 5 shows the parameter estimates of a linear model for the difference in HbA1c, produced by the code below.

```
proc glmselect data=final;
   class annual_income education_level Master_diag sex group white
   black(ref='0') hispanic(ref='0')/param=ref;
   model a1cdiff=Age_Enroll annual_income education_level Master_diag sex
   group white black hispanic/selection=backward(select=SL sls=.15)
   showpvalues;
run;
```

Parameter Estimates						
Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-0.370241	0.334513	-1.11	0.2739
income	Over $15,000	1	-0.366586	0.657998	-0.56	0.5800
income	Under $15,000	1	0.744560	0.420557	1.77	0.0830
black	1	1	-2.479037	0.443622	-5.59	<.0001

Figure 5: Parameter estimates from model of difference in HbA1c measure

The annual income of participants is a significant predictor in their estimated change in HbA1c level. Those that fall into the over $15,000 per year category experience the largest decrease in HbA1c, with a reduction of .37 percentage points compared to the reference category of those that declined to provide their income. Black participants experienced an estimated 2.5 percentage point decrease in HbA1c measure compared to non-black participants.

The PAM-13 assessment was given to participants at enrollment and every three months thereafter. A higher score indicates more confidence and knowledge from the patient in self-managing their condition; therefore, a larger increase is desirable. Figures 6 shows the parameter estimates of a linear model for the difference in PAM-13 score, produced by the code below.

```
proc glmselect data=final;
   class annual_income education_level Master_diag sex group white
   black(ref='0') hispanic(ref='0')/param=ref;
   model pamdiff=Age_Enroll annual_income education_level Master_diag sex
   group white black hispanic/selection=backward(select=SL sls=.15)
   showpvalues;
run;
```

Parameter Estimates						
Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-70.239119	32.733647	-2.15	0.0575
Age_Enroll		1	1.267836	0.615922	2.06	0.0666
income	Over $15,000	1	12.147932	12.961875	0.94	0.3707
income	Under $15,000	1	-24.946827	9.494520	-2.63	0.0253
Master_diag	1	1	9.700830	10.175112	0.95	0.3629
Master_diag	2	1	27.314482	9.032368	3.02	0.0128
black	1	1	26.818654	9.285526	2.89	0.0162

Figure 6: Parameter estimates from model of difference in PAM-13 score

Age, income, diagnosis, and race are all significant predictors in modeling participants' change in PAM-13 score. For every one year increase in age, the estimated difference in PAM-13 score increases by 1.3. Participants with an income over $15,000 per year experience an estimated 12.1 point increase in score, while those with an income under $15,000 have an estimated decrease of 24.9 points, both compared to the reference group of those that declined to provide their income. Heart failure patients have a predicted 9.7 point decrease in their score, while diabetes patients have a predicted 27.3 point increase in their score, both compared to the reference group of participants diagnosed with both diseases. Black participants experience an estimated 26.8 point increase in score, compared to those that are non-black.

Due to demographic variables being highly correlated and dependent on one another, each of these claims can only be made holding all other factors constant.

Using the code below, a logistic model was built with inpatient hospital utilization as the response, and produced the output in Figure 7.

```
proc glmselect data=final;
   class annual_income education_level Master_diag sex group white
   black(ref='0') hispanic(ref='0')/param=ref;
   model hosp=Age_Enroll annual_income education_level Master_diag sex
   group white black hispanic/selection=backward(select=SL sls=.15)
   showpvalues;
run;
```

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.050877	0.090924	-0.56	0.5780
education College	1	0.338986	0.121025	2.80	0.0069
education Grade 12 or GED (High School Graduate)	1	0.179020	0.119136	1.50	0.1384
education Grade School	1	0.090409	0.122028	0.74	0.4618
group Control	1	0.135672	0.091874	1.48	0.1453

Figure 7: Parameter estimates from logistic model of hospital utilization

Participants' education level and treatment group were significant predictors in estimating whether they would have a hospital visit during the clinical trial. Those with any college experience have the largest increase in their log-odds of having a hospital stay with a coefficient of .34, or a $e^{.34}$ =1.40 multiplicative increase in the probability of having an inpatient hospital stay. Participants assigned to the control group experienced an estimated .14 multiplicative increase in the log-odds, or a 1.15 multiplicative increase in the probability of having an inpatient hospital stay.

A logistic model with emergency department utilization as the response was built. Backward selection was performed using the same methodology as above, but all main effects dropped out of the model. This suggests that patients' emergency department visits are not associated with any demographic factors.

BUILDING A LOG-LINEAR MODEL

The continuous response variables were categorized dichotomously, and log-linear models were built using PROC CATMOD. A decrease of at least one percentage point in HbA1c is considered to greatly improve a diabetic person's management of their disease. Participants were categorized as having at least a one point decrease in their HbA1c measure over the duration of the clinical trial, or not having a one point decrease. This new variable was then used to build a log-linear model with independent demographic variables of interest. The following code produces a log-linear model to predict whether a participant had a decrease in their HbA1c measure of at least one point, and the output is shown in Figures 8 and 9.

```
proc catmod data=a1c order=data;
   model over_one*group=_response_ race*_response_ /pred param=effect;
   loglin over_one group;
run;
```

Maximum Likelihood Analysis of Variance			
Source	DF	Chi-Square	Pr > ChiSq
over_one	1	1.96	0.1620
group	1	5.30	0.0213
Race*over_one	3	11.87	0.0078
Race*group	3	1.56	0.6693
Likelihood Ratio	4	5.80	0.2146

Figure 8: Maximum likelihood ANOVA from log-linear model of binary HbA1c

Analysis of Maximum Likelihood Estimates					
Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
over_one	1	-0.2143	0.1533	1.96	0.1620
group	FoodRx	0.3299	0.1432	5.30	0.0213
Race*over_one	Black 1	0.8037	0.2537	10.04	0.0015
	Hispanic 1	-0.1722	0.2323	0.55	0.4583
	Other 1	-0.0411	0.3003	0.02	0.8912
Race*group	Black FoodRx	0.1079	0.2365	0.21	0.6484
	Hispanic FoodRx	0.1849	0.2333	0.63	0.4280
	Other FoodRx	-0.0745	0.2953	0.06	0.8009

Figure 9: Maximum likelihood estimates from log-linear model of binary HbA1c

The log-linear model built consists of a two by two table of dependent variables HbA1c reduction over one and treatment group. This table is then adjusted for by the independent demographic variables. In this case, only race was significant as seen in the ANOVA table in Figure 8. The likelihood ratio is insignificant, indicating a good model fit. The two-way interaction of race and over one point reduction is significant, indicating an association among those variables.

The difference in PAM-13 scores from the beginning to the end of the program was dichotomously classified into an increase in score, or decrease or no change. Only three participants experienced an increase in their PAM-13 score over the duration of the clinical trial. This is partially due to not everyone having more than one measure. This lack of data prevented a meaningful log-linear model for PAM-13 from being fit.

The difference in daily servings of fruit and vegetables was dichotomously classified into an increase in servings, or a decrease or no change in servings. This variable was used as the response in a log-linear model using the code below, and the output is shown in Figures 10 and 11.

```
proc catmod data=fruit order=data;
   model fruit*group=_response_ sex*_response_ /pred param=effect;
   loglin fruit group;
run;
```

Maximum Likelihood Analysis of Variance			
Source	DF	Chi-Square	Pr > ChiSq
fruit	1	4.36	0.0368
group	1	5.68	0.0171
Sex*fruit	1	2.87	0.0905
Sex*group	1	0.21	0.6489
Likelihood Ratio	2	2.22	0.3303

Figure 10: Maximum likelihood ANOVA from log-linear model of binary fruit and vegetable servings

Analysis of Maximum Likelihood Estimates					
Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
fruit	1	-0.3116	0.1493	4.36	0.0368
group	Control	-0.3675	0.1542	5.68	0.0171
Sex*fruit	Female 1	-0.2527	0.1493	2.87	0.0905
Sex*group	Female Control	0.0702	0.1542	0.21	0.6489

Figure 11: Maximum likelihood estimates from log-linear model of binary fruit and vegetable servings

The likelihood ratio is insignificant indicating a good model fit, and the two-way interaction between the sex of the participant and dichotomous fruit variable is significant, indicating an association between those variables. The coefficient for the interaction between sex and an increase in fruit and vegetable servings is -.25, indicating a multiplicative decrease of .78 in the probability of a female having an increase in servings compared to males.

The difference in physical activity sessions per week was dichotomously classified into an increase in sessions, or a decrease or no change in session. A log-linear model was built using this variable as the response, but no demographic variables were significant predictors. Two log-linear models were also built using emergency department and hospital utilization as response variables. None of the demographics as independent variables was significant in either model. This is consistent with the logistic model for emergency department utilization, where all terms also dropped out of the model.

COMPARING RESULTS OF DIFFERENT METHODS

The linear and logistic regression models of the six response variables of interest always had more significant predictors than the log-linear models for the same response. Both models for HbA1c are the only instance where there is a predictor in common between the variables. The indicator variable for black participants in the linear model shows a decrease in their HbA1c measure compared to non-black participants. The log-linear model for HbA1c has a positive coefficient of .80 for black participants, indicating an estimated 2.23 multiplicative increase in the probability of a black participant experiencing an HbA1c decrease of over one percentage point, compared to the reference group of white participants.

CONCLUSION

The results of the multiple linear regression and logistic models show that the treatment group to which the participant was assigned is not a significant predictor in all except one of the models. This suggests that socioeconomic and demographic factors are more influential on health and lifestyle choices than purely having access to fresh fruits and vegetables. Due to a relatively small sample size, results from this analysis should be taken with caution. A more liberal significance level of .15 was used because the intent with these models was to predict human behavior, and for the most part humans are very unpredictable. Biases other than those previously discussed are also present. Participants were selected

from those already enrolled in a program at Healthier Communities, and may not be representative of the population of Grand Rapids. The control group was approximately half the size of the treatment group and this can mostly be attributed to people being more likely to participate if they were going to receive the food vouchers. This clinical trial gave Healthier Communities an opportunity to educate low-income individuals on the importance of a diet rich in fresh fruits and vegetables, and the influence it has on your health.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Shana Kelly
Spectrum Health: Healthier Communities
665 Seward Ave NW, Suite 110
Grand Rapids, MI 49504
Shana.kelly@spectrumhealth.org