Paper 3282-2019

Integrating Case Studies in a Health Analytics Curriculum

Tyler C Smith, MS, PhD, National University and Besa Smith, MPH, PhD, ICON, plc

ABSTRACT

Advanced informatics layered with customized analytics has driven major change in every discipline in the past decade. In the healthcare environment, individual level determinants of health are being leveraged to reveal personalized patient management that considers disease patterns, high-risk attributes, hospital acquired conditions, and performance measures for specialized treatment approaches. In light of these health informatics advancements during the last decade, the growing need for health analytics expertise has created a critical void in higher education. An abundance of new data-based opportunities that have made large public-use data sets accessible for easy download and use in the classroom have allowed for a much more applied classroom experience. More hands-on applications are positioning students for greater impact in the real world upon graduation and entry into the job market. In this presentation, we highlight the use of controlled and adaptive case studies leveraging SAS® and real-world data to provide a more realistic classroom experience.

INTRODUCTION

Informatics solutions, data storage advances, and increased data accessibility offer unique opportunities in the classroom to study health related data. This paper presents analytic case studies using publicly available health datasets and describes various limitations and strengths of these data pertaining to the use in the classroom as well as the publication of analyses using these types of data.

PUBLICLY AVAILABLE HEALTH RELATED DATA

Over the past decade, many health-related data sets have been made publicly available for analyses, complementing ongoing investigations within the original scope of the data collection. Strict management of these data prepared by investigators or data suppliers with the intent of de-identifying and making them available for public use allows researchers to obtain these data without special permissions. Institutional Review Boards (IRB) or other entities whose main objective is the protection of human subjects in research have acknowledged that the analysis of de-identified, publicly available data do not constitute human subjects research as defined by 45 CFR 46.102.

There are many advantages to using these data including both time and money. The readily available datasets are often large, have many variables to use as endpoints, variables of interest,

and measurable confounders. Further, many data sets have accumulated serial cross sections over many years of data collection and present a reasonable representativeness of these sampled populations. Most datasets also include variables for weighting based on the inverse of the sampling scheme and inverse of the response patterns allowing for enhanced population estimates of the sampling frame and error terms in estimation. These datasets are also accompanied by detailed code books (data dictionaries) and some include sample programming code.

Using these data for case studies in a classroom setting is acceptable without the oversight of an Institutional Review Board (IRB) (check with your specific IRB for guidance on how they may handle public use data). If these data, however, are going to be used for capstone or a thesis to be published, a determination letter should be sought by the academic institution's IRB.

Of the greatest disadvantage in using publicly available datasets is that the data were most likely collected for different reasons than the intent of your research and may not meet the specific objectives of what you wish to investigate had you conducted primary research collection. Additionally, the sampling frame may not include your targeted population nor may the variables be measured ideally toward the hypothesis you wish to test. Further, limitations due to sample size, response rate, or lack of information on confounding need to be considered. Finally, these data are de-identified and though we may wish to link elsewhere, we cannot. Despite these limitations, these data can provide excellent opportunities for hypothesis testing prior to large financial investment, with analyses completed in a relatively short period of time. In the classroom, they offer a real-life view of the limitations and strengths of data in general and offer additional areas for development of the student.

FORWARD TO CASE STUDIES

Case studies present examples of real-life health analytic scenarios that allow for hands-on exploring of a relevant issue. Complementary to textbooks and traditional approaches, case studies tap different parts of the brain and provide for an experience closer to what students will find in the real-world when they leverage their tools to conduct an analysis followed by an interpretation of the results. Case studies may be short, which often work effectively for specific statistical tests or analytic concepts, or they may be longer when a continuation of a full-analysis is sought. Case studies are thought provoking and often excite a student by the prospect of analyzing a health topic relevant to society or their own personal interests. The best case studies are those that may have multiple "correct" answers or which yield varying results depending upon the methods chosen to address the objective. Analytical case studies need to be concise though a review of the literature for context should be required.

After completing several class meeting-based, group-based, or individual-based case studies, it is very helpful to flip the approach and ask the students to draft their own case study. This sets a strong foundation for the process and allows the student to consider whether the objective is framed with enough information and whether the data exist to answer the objective. These can be fun as well as educational, providing an opportunity for the student to think through the

process well enough to draft their own case study. The case studies can then be randomly assigned to teams, each competing to complete the analysis and then assembling back as a class to interpret findings. Allowing for interpretation/reasoning for both the case study formulation team as well as the case study investigation team allows for rich discussion.

Many of the critical elements of framing an objective, building an analytic data set, running the analysis, and interpreting the results can all lead to fascinating educational moments that instructors may not anticipate. The elucidation of a previously unknown option or proc during these student-generated investigations reminds us all that there are many different ways to bake a cake. Allowing assessment of the case study, assessment of the analysis and interpretation offer even further understanding to the students.

Remember that it is perfectly fine to run into an approach or an issue for which you do not have the answer. These times provide another learning opportunity to review SAS Global or SAS Regional white papers that have a tremendous amount of "how to" at the student's finger tips. Sometimes the process of learning an obscure option or macro is as important as the option or macro—embrace the journey!

IN GENERAL

How many times has a student triumphantly handed over a regression output and asked where to find the p-value? One of the hardest elements to put forth is the idea of patience and completing the necessary steps before the regression is run. To that end…there are many…many steps to go through before we can hit "run" and here are a few of the standard necessities of any analysis (there are many others and each analyst has their flavor of what they like to review).

First, don't underestimate how difficult it is to frame an objective or draft an appropriate null hypothesis. Take some time with this and show many examples!

For instance, which of these would you consider to be better?

1) The objective of this analysis is to investigate mental health.

2) The objective of this analysis is to investigate differences in hypertension in those who consume large quantities of salt when compared to those who do not.

3) The objective of this analysis is to investigate differences in BMI in those with and without Type 2 Diabetes after controlling for age, gender, physical activity, and marital status in the 2016 BRFSS data.

The first one will clearly take the student the rest of their collegiate life to investigate, the second presents an idea but leaves us wondering a little too much. The third presents an objective, confounders, and a data set.

After a clean objective is framed and confirmation that the data and variables to address the objective are available, a model statement should be written. Not a programming model

statement or a mathematical model statement, but a theoretical model statement with the variables, the variable types, and how data will be coded (categorical or continuous). Including this in the program and just prior to the data step for the creation of the analytic data set is an excellent place to locate this road map.

*The objective of this study is to investigate the association between Diabetes and CVD after controlling for gender and physical activity in BRFSS 2014 respondents 65 or older;

* Y diabetes (1=yes/ 3=no) = X1 CVD (1=yes/ 2=no) + X2 sex (1=male/2=female) + X3 activity (1=yes and 2=no) ;

Remember…75% of your time will be spent on the data management, plan for it, embrace it, you will be far better off if you take the extra time and care and create a well thought out analytic data set. Make sure you spend this amount of time with your students working out the complexities of the data set.

CASE STUDY

The National Healthcare Discharge Survey (NHDS) will be used for the first case study (https://www.cdc.gov/nchs/nhds/index.htm). Though most analytic tool sets are portable to many disciplines, approaches specific to a discipline and understanding data specific to a discipline is critical in the interpretation phase of an investigation. The NHDS allows for the investigation of ICD-9 discharge codes which are a mainstay of health-related outcomes research.

- Characteristics of inpatients discharged from non-Federal short-stay hospitals in the United States
- Conducted annually (1965 to 2010)
- NHDS encompasses patients discharged from noninstitutional hospitals, exclusive of military and Department of Veterans Affairs hospitals, located in the 50 States and the District of Columbia
- National Hospital Care Survey (NHCS) integrates NHDS with the emergency department (ED), outpatient department (OPD), and ambulatory surgery center (ASC) data collected by the National Hospital Ambulatory Medical Care Survey (NHAMCS).
- Very large with approximately 150,000 patients per year
- Codebook and survey available
- Many peer-reviewed papers as well as reports written based on these data
- Information about limitations and strengths included

Though these secondary data analyses of existing publicly available data are technically not required to have IRB oversight, it is best to require that the students obtain human subjects training. CITI is a great program and required certificates early in the program give students a robust understanding of human subjects' research.

From this link (https://www.cdc.gov/nchs/nhds/index.htm), find the "Questionnaires, Datasets, and Related Documentation" tab on the left and enter. Find the two links: "NHDS Downloadable Data Files via FTP" and "NHDS Downloadable Documentation via FTP". Enter each link and download the "NHDS08", "NHDS09", "NHDS10" data (need to ftp it with link at top) and the documentation for each "NHDS_2010_Documentation.pdf".

Please consider these numbers.

- Asthma is estimated to account for one-quarter of all emergency room visits in the U.S. each year, with 1.8 million emergency room visits.
- Each year, asthma accounts for more than 10 million outpatient visits and 479,000 hospitalizations.
- The average length of stay (LOS) for asthma hospitalizations is 3.6 to 4.3 days.
- Nearly half (44%) of all asthma hospitalizations are for children.

The problem: You are conducting a study of asthma and geography and wish to investigate whether there is an association between asthma diagnosis and region.

You believe that conducting an adjusted analysis is appropriate though you are concerned that other factors need to be matched on or otherwise taken into account differently than in an adjusted analysis. You feel that the propensity of "treatment" (geographic) assignment is conditional on at least two observed baseline characteristics: 1) Hospital Ownership, 2) Type of Admission

QUESTION FOR STUDENT: Write the objective

The objective of this analysis is to investigate the association between asthma diagnosis and geographic region after adjusting for age, gender, and race, while also accounting for hospital ownership and type of admission associations with geography in the NHDS 2008-2010 data.

QUESTION FOR STUDENT: Write the model statement

Asthma diagnosis (yes,no) = geographic region (Northeast, Midwest, South, West) + age (18-40, 41-60, >60) + gender (male, female) + race (White, Black, Asian, Other)

From the model statement and the well-thought out plan for the variables, the student will know if they are going to be working with categorical data (pathway towards chi-square, logistic regression, and other categorical data methods) or continuous data (pathway towards t-tests, correlations, ANOVA, GLM, and other continuous methods). This paper will focus on a categorical data approach though it is simple to go to a continuous case study approach by changing the outcome variable to continuous.

Goals for this Case Study: Restrict the population to 18 and older and conduct an unadjusted analysis for "table 1" (by exposure) and "table 2" (by outcome) investigations. Then continue with the analysis by adjusting for age, gender, and race in an adjusted analysis for "table 3". Lastly, use propensity scores for the two factors (hospital ownership and type of admission) and selection by region to adjust your results. Please present in table form and interpret.

Reading in data, remember flat files?

*These input statements read in columns of your flat file and name the column(s) anything you wish. Resist calling the variables x1-xn as these will be more confusing later and instead take the data layout and name the variables close to what they are.

```
data NHDS10;  *read in NHDS 2010 data.
infile 'C:\ PATHWAY TO YOUR DATA \NHDS10.pu.txt';
  input  surveyyear 1-2   Newborn 3   Ageunits 4   ageyears 5-6   sex 7   race 8   marital 9
       dischargeMonth 10-11   dischargestatus 12   dayscare 13-16   LOS 17   region 18
       numbbeds 19   hospowner 20   Analysisweight 21-25   twodigitssurveyyear 26-27
       dx1 $ 28-32   dx2 $ 33-37   dx3 $ 38-42   dx4 $ 43-47  dx5 $ 48-52   dx6 $ 53-57   dx7
       $ 58-62  dx8 $ 63-67   dx9 $ 68-72   dx10 $ 73-77   dx11 $ 78-82   dx12 $ 83-87   dx13
       $ 88-92   dx14 $ 93-97   dx15 $ 98-102   proc1 $ 103-106   proc2 $ 107-110   proc3
       $ 111-114   proc4 $ 115-118  proc5 $ 119-122   proc6 $ 123-126   proc7 $ 127-130
       proc8 $ 131-134   prisourcepayment 135-136  secrourcepayment  137-138   drg 139-
       141   admisstype 142  admisssource 143-144  admissdxs $ 145-149;
run;

data NHDS09;  *read in NHDS 2009 data.
infile 'C:\PATHWAY TO YOUR DATA\NHDS09.pu.txt';
  input  surveyyear 1-2   Newborn 3   Ageunits 4   ageyears 5-6   sex 7   race 8   marital 9
       dischargeMonth 10-11   dischargestatus 12   dayscare 13-16   LOS 17   region 18
       numbbeds 19   hospowner 20   Analysisweight 21-25   twodigitssurveyyear 26-27
       dx1 $ 28-32   dx2 $ 33-37   dx3 $ 38-42   dx4 $ 43-47  dx5 $ 48-52   dx6 $ 53-57   dx7
       $ 58-62  proc1 $ 63-66   proc2 $ 67-70   proc3 $ 71-74   proc4 $ 75-78
       prisourcepayment 79-80  secrourcepayment  81-82  drg 83-85   admisstype 86
       admisssource 87-88  admissdxs $ 89-93;
run;
```

```
data NHDS08; *read in NHDS 2008 data.
infile 'C:\ PATHWAY TO YOUR DATA \NHDS08.pu.txt';
  input  surveyyear 1-2   Newborn 3   Ageunits 4   ageyears 5-6   sex 7   race 8   marital 9
       dischargeMonth 10-11   dischargestatus 12   dayscare 13-16   LOS 17   region 18
       numbbeds 19   hospowner 20   Analysisweight 21-25   twodigitssurveyyear 26-27
       dx1 $ 28-32   dx2 $ 33-37   dx3 $ 38-42   dx4 $ 43-47  dx5 $ 48-52   dx6 $ 53-57   dx7
       $ 58-62
       proc1 $ 63-66   proc2 $ 67-70   proc3 $ 71-74   proc4 $ 75-78  prisourcepayment 79-
       80  secrourcepayment  81-82   drg 83-85   admisstype 86  admisssource 87-88
       admissdxs $ 89-93;
run;
```

Notes:

1) The variables that are the same over the 3 data sets are given the same variable name. If they are not, when concatenated, there will be different variables that will be completely missing for that dataset.

2) There are additional variables and procedures in the 2010 data that need to be accounted for.

3) If any of these variables are categorized differently, they will not integrate. For example, if the data managers labeled categories of gender as "M" and "F" for one year and "1" and "2" for another year, you will have 4 categories for your gender variable after these data are appended.

*Next, we need to make sure the 3 data sets will align with variable names.

```
data NHDS10; *drop variables out of this data set that do not exist in the other 2 data sets.
    This is also   where you would recategorize variables if needed to make sure all 3 data
    sets line up.
    set NHDS.NHDS10 (drop= dx8 dx9 dx10 dx11 dx12 dx13 dx14 dx15 proc5 proc6 proc7
    proc8);
run;

*Now append the datasets for the years 2008, 2009, 2010; ** The ending data set should
contain all 3 data sets;

proc append base=NHDS08 data=NHDS09;*the NHDS08 data now has NHDS09
concatenated onto it.
run;
proc append base=NHDS08 data=NHDS10a;  *the new NHDS08 data now has NHDS10
concatenated onto it.
run;
```

QUESTION FOR STUDENT: Read David Carr's paper on PROC APPEND, are there other methods to concatenate data sets? Why use PROC APPEND? (While the current paper is not about PROC APPEND Carr's paper and others that will help the discussion and growth of the students.)

proc contents data=nhds08; *Note that it is called NHDS08. **QUESTION:** Confirm the number of observations and number of variables;
run;

```
111  *Now append the datasets for the years 2008, 2009, 2010;  ** The ending data set should
111! contain all 3 and call it nhds200820092010;
112  proc append base=NHDS08 data=NHDS09;
113   run;

NOTE: Appending WORK.NHDS09 to WORK.NHDS08.
NOTE: There were 162151 observations read from the data set WORK.NHDS09.
NOTE: 162151 observations added.
NOTE: The data set WORK.NHDS08 has 327781 observations and 33 variables.
NOTE: PROCEDURE APPEND used (Total process time):
      real time           0.07 seconds
      cpu time            0.07 seconds


114  proc append base=NHDS08 data=NHDS10a;
115   run;

NOTE: Appending WORK.NHDS10A to WORK.NHDS08.
NOTE: There were 151551 observations read from the data set WORK.NHDS10A.
NOTE: 151551 observations added.
NOTE: The data set WORK.NHDS08 has 479332 observations and 33 variables.
NOTE: PROCEDURE APPEND used (Total process time):
      real time           0.07 seconds
      cpu time            0.07 seconds
```

Notice that we have 479,332 and 33 variables.

Use an array to scan the diagnosis codes for Asthma.
QUESTION FOR STUDENT: First, go to this link to look up asthma and determine the ICD-9 diagnostic code (https://www.findacode.com/icd-9/icd-9-cm-diagnosis-codes.html).

Build your analytic data set using your model statement as a pathway

Asthma diagnosis (yes,no) = geographic region (Northeast, Midwest, South, West) + age (18-40, 41-60, >60) + gender (male, female) + race (White, Black, Asian, Other);

```
Data nhds200820092010;
  set NHDS08 (where= (  (admisstype in (1,2,3)) and (ageyears>=18) and (race in (1,2,3,4,5,6,8))
));
          *the where statement allows us to restrict to 18 and over as well as admission type 1,2,3
and race in 1,2,3,4,5,6;

asthma=0;
 array d(7) dx1-dx7;
```

```
      do i=1 to 7;
        if (substr(d(i),1,3) in ( '493' )) then asthma=1;
      end;

  **categorize the continuous age variable into a new variable called agecat with 3 levels;
  if 18<=ageyears<=40 then agecat=1;
  if 41<=ageyears<=60 then agecat=2;
  if 61<=ageyears then agecat=3;


  **recategorize race into a 4-level variable;
  racecat=0;
  if race in (1) then racecat=1;
  if race in (2) then racecat=2;
  if race in (4) then racecat=3;

run;
*this is a complete case analysis approach.  QUESTION: How many observations were lost?;

proc freq data=nhds200820092010;
  tables asthma;
run;
```

QUESTION FOR STUDENT: What is the percentage of asthma diagnoses? 5.4%

QUESTION FOR STUDENT: Read the papers by Waller and Kuligowski, are there other ways to find the asthma diagnosis other than using an array?

Continuous versus Discrete

Categorizing the age in years variable presents a great opportunity to discuss the reasons for and against categorizing. See the paper by David Pasta for an interesting read and great discussion points.

```
*Formats are an important output benefit, see the paper by Shoemaker for more information on
PROC FORMAT;

proc format; *this will allow you to label your output with meaningful words or values;
  value yn_fmt   0='No' 1='Yes';
  value geo_fmt  1='Northeast' 2='Midwest' 3='South' 4='West';
  value admtypfm 1 = 'Emergency' 2 = 'Urgent' 3 = 'Elective' ;
  value hospown  1='Proprietary' 2='Government' 3='Nonprofit, including church';
  value sex_fmt  1='Male' 2='Female';
  value age_fmt  1='18-40' 2='41-60' 3='61 or older';
  value racefmt  1='White' 2='Black/AfAmerican' 3='Asian' 4='Other';
run;
```

Using the format and a PROC FREQ, we can get the confounders crossed by the variable of interest "region". Note, the "chisq" requests chi-square statistics be given in the output.

The format statement labels the output based on above proc format.

```
proc freq data=nhds200820092010; *table 1 are variables crossed by exposure (variable) of interest;
  tables (agecat sex racecat) * region / chisq;
  format  region geo_fmt. agecat age_fmt. sex sex_fmt. racecat racefmt. ;
run;
```

Frequency
Percent
Row Pct
Col Pct

Table of sex by region

sex	region				
	Northeast	Midwest	South	West	Total
Male	27000 8.76 21.62 42.15	27172 8.82 21.76 39.87	58890 19.11 47.16 39.99	11815 3.83 9.46 41.07	124877 40.51
Female	37064 12.02 20.21 57.85	40979 13.29 22.35 60.13	88373 28.67 48.19 60.01	16950 5.50 9.24 58.93	183366 59.49
Total	64064 20.78	68151 22.11	147263 47.77	28765 9.33	308243 100.00

Statistics for Table of sex by region

Statistic	DF	Value	Prob
Chi-Square	3	103.0071	<.0001

This is a perfect place to spend time discussing the Chi-Square statistic (103.007), what the p-value indicates (p-value<0.0001), and where the proportional differences are.

Chi-square statistic formula:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

QUESTION FOR STUDENT: With a p-value of <0.05, what does the chi-square test tell us?

QUESTION FOR STUDENT: What was the proportion of females? 59.5%.

QUESTION FOR STUDENT: What was the proportion of those from the south? 47.8%.

QUESTION FOR STUDENT: Among males, what proportion is from the west? Using the row percent, 9.5%.

QUESTION FOR STUDENT: Among those living in the northeast, what proportion is female? Using the column percent, 57.9%

Spend time going over the cross tabs for the confounders by the exposure/variable of interest.

TABLE 2

```
proc freq data=nhds200820092010; *table 2 are unadjusted analyses of variables crossed by outcome asthma diagnosis;
  tables (region agecat sex racecat) * asthma / chisq;
  format  region geo_fmt. agecat age_fmt. sex sex_fmt. racecat racefmt. asthma yn_fmt.;
run;
```

SAS Output for region by asthma:

Frequency Percent Row Pct Col Pct	Table of region by asthma		
	asthma		
region	No	Yes	Total
Northeast	60109 19.50 93.83 20.61	3955 1.28 6.17 23.79	64064 20.78
Midwest	64582 20.95 94.76 22.15	3569 1.16 5.24 21.47	68151 22.11
South	139617 45.29 94.81 47.88	7646 2.48 5.19 46.00	147263 47.77
West	27312 8.86 94.95 9.37	1453 0.47 5.05 8.74	28765 9.33
Total	291620 94.61	16623 5.39	308243 100.00

Statistics for Table of region by asthma

Statistic	DF	Value	Prob
Chi-Square	3	97.9869	<.0001

QUESTION FOR STUDENT: How would these numbers be presented in a table?

TABLE 2. Univariate Associations of Characteristics of 308,243 NHDS 2008-2010 Patients By Asthma Diagnosis.

Variable	Population *N*(%)		No Asthma DXS *n*(%) (N=291,620)		Asthma DXS *n*(%) (N=16,623)		*p* value*
Region							
Northeast	64,064	(20.8)	60,109	(20.6)	3,955	(23.8)	
Northwest	68,151	(22.1)	64,582	(22.2)	3,569	(21.5)	
South	147,263	(47.8)	139,617	(47.9)	7,646	(46.0)	
West	28,765	(9.3)	27,312	(9.4)	1,453	(8.7)	<.0001
Age Category							
18-40	71,202	(23.1)	66,855	(22.9)	4,347	(26.2)	
41-60	84,126	(27.3)	78,072	(26.8)	6,054	(36.4)	
61 or older	152,915	(49.6)	146,693	(50.3)	6,222	(37.4)	<.0001
Sex							
Male	124,877	(40.5)	120,215	(41.2)	4,662	(28.1)	
Female	183,366	(59.5)	171,405	(58.5)	11,961	(71.9)	<.0001
Race							
White	232,186	(75.3)	220,724	(75.7)	1,046	(69.0)	
Black or	52,610	(17.1)	48,656	(16.7)	11,462	(23.8)	
Asian	4,476	(1.5)	4,315	(1.5)	3,954	(1.0)	
Other	18,971	(6.2)	17,925	(6.2)	161	(6.3)	<.0001

* *p* values based on Pearson chi-square test of association.

RESULTS: Table 2 presents the unadjusted associations between asthma and region along with age, sex, and race. There were 16,623 (5.4%) of the population with a discharge diagnosis of asthma. There were proportionately more then expected based on the overall population from the Northeast, younger, female and Black or African Americans who had an asthma diagnosis (p-values for all variables <0.0001).

This presents a good time to discuss clinical versus statistical significance as well as statistical significance when investigating large data.

ADJUSTED ANALYSES USING PROC LOGISTIC

Logistic regression is a statistical method used to evaluate many independent variables (X_1, X_2, …, X_p) in order to predict a dichotomous outcome. Generally this outcome is denoted as Y = 1 or Y = 0 for the two possibilities.

In logistic regression the probability of an occurrence of the outcome being investigated is defined as:

$$P(Y=1) = \frac{1}{1+\exp[-\beta_0 + (\sum_{k=1}^{p} \beta_k X_k)]}$$

SAS offers PROC LOGISTIC which is a procedure for fitting regression models for binary or ordinal outcomes.

```
proc logistic data = nhds200820092010;
class   asthma (ref='0')  region (ref='1')  agecat (ref='1') sex (ref='1') racecat (ref='1') / param=ref;
model asthma = region agecat sex racecat;
run;
```

Data= nhds200820092010 names the input data set for the logistic regression.

Class statement allows us to establish the reference category in the categorical variables without first making "dummy" variables in a data step. In this case, we are using reference cell coding.

Param=reference requests that the parameter estimates, odds ratios, and confidence intervals be calculated using reference cell coding. The default parameter estimates would be computed using the effect coding scheme which estimates the difference in the effect of each non-reference level compared to the average effect over the other levels of the variable.

Model= requests that Asthma is the end point (we are modeling the probability of Asthma (yes)).

This is where you can go into stepwise regression, standardized coefficients, Hosmer Lemeshow, c-statistics, and much more.

There are **MANY** options that are not discussed here and can be found at:
https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect016.htm

Adjusted, Non-weighted SAS Output:

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
region	3	149.9041	<.0001
agecat	2	1017.2208	<.0001
sex	1	1119.0632	<.0001
racecat	3	408.6616	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
region 2 vs 1	0.821	0.783	0.860
region 3 vs 1	0.783	0.752	0.815
region 4 vs 1	0.810	0.761	0.862
agecat 2 vs 1	1.345	1.291	1.402
agecat 3 vs 1	0.739	0.709	0.769
sex 1 vs 2	0.548	0.529	0.568
racecat 0 vs 1	1.063	0.995	1.136
racecat 2 vs 1	1.450	1.396	1.507
racecat 3 vs 1	0.637	0.543	0.747

RESULTS: All variables were statistically significant at the alpha=0.05 level. After controlling for age, sex, and race, those residing in the south were at 0.78 times the odds of having an asthma diagnosis when compared to those living in the northeast (OR=0.78; 95% CI = 0.75, 0.82).

WEIGHTING

Data are often collected with complex sampling designs to ensure subgroup representation and other statistical and methodological efficiencies. There are often response differences across subgroups as well. Data should be weighted if the sample design does not give each individual an equal chance of being selected or when certain subgroups have differing probabilities of response. For example, households which have equal selection probabilities but one person is interviewed from within each household result in people from large households having a smaller

chance of being interviewed. Weights are designed to lessen or eliminate the burden of sampling or response issues.

See the papers by the following authors that will give you a great appreciation for weighting: Smith; Lohr; Berglund; Lewis; and Cassell

For more information on PROC SURVEYLOGISTIC, visit the SAS Support Site:
https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_surveylogistic_sect001.htm

https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_surveylogistic_a0000000337.htm

PROPENSITY SCORES

Suppose that you want to evaluate the effect of an intervention though you do not randomly allocate at entry into the study. You have data on every participant, race, ethnicity, marital status, family income, age, gender, etc. and you feel that the selection of group membership for the intervention group/control is informed by the characteristics of the participants.

This can be a real issue with secondary data analyses in healthcare. Propensity scores may help by matching or adjusting out some of this informative selection.

The propensity score is the probability of treatment assignment conditional on observed baseline characteristics. In theory…the propensity score may allow one to design and analyze an observational (nonrandomized) study so that it mimics some of the particular characteristics of a randomized controlled trial. The idea is to compare individuals who, based on observables, have a very similar probability of receiving treatment (similar propensity score), but one of them received treatment and the other did not. The research may choose to match on the propensity score or use a covariate adjustment leveraging the propensity score.

In this case, we will create the propensity scores based upon hospital ownership and admission type using the following PROC LOGISTIC. Here, the endpoint is region.

```
***now output the propensity for group selection by indicating region as the y variable and
regressing hospital ownership and admission type onto region;
proc logistic data = nhds200820092010;
class   region (ref='1') hospowner (ref='1')  admisstype (ref='1')   / param=ref;
model region = hospowner admisstype;
OUTPUT OUT=AllPropen prob=prob; *Output the propensity for group selection (that is region
selection);
run;
```

"OUT=AllPropen prob=prob" will output a data set with a new variable named "prob". This is the variable that you now included in your previous logistic regression.

```
proc logistic data = AllPropen;
class   asthma (ref='0')  region (ref='1')  agecat (ref='1') sex (ref='1') racecat (ref='1') / param=ref;
model asthma = region agecat sex racecat prob;
run;
```

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
region 2 vs 1	0.821	0.799	0.843
region 3 vs 1	0.782	0.764	0.800
region 4 vs 1	0.809	0.781	0.839
agecat 2 vs 1	1.345	1.314	1.378
agecat 3 vs 1	0.739	0.722	0.756
sex 2 vs 1	1.824	1.787	1.861
racecat 0 vs 1	1.062	1.022	1.104
racecat 2 vs 1	1.450	1.418	1.483
racecat 3 vs 1	0.637	0.581	0.698
prob	0.984	0.954	1.014

Though there was not a lot of propensity score adjustment to the final adjusted ORs and CIs, there was enough to make for interesting conversation and hopefully see the utility of propensity scores when using large data warehouses that are now being maintained across healthcare.

OTHER EXCELLENT HEALTH RELATED SERIAL CROSS-SECTIONAL PUBLICLY AVAILABLE DATA

Behavioral Risk Factor Surveillance System (BRFSS) is a health survey which evaluates behavioral risk factors and chronic diseases. It is administered by the Centers for Disease Control and Prevention and conducted by individual state health departments. The survey is the world's largest telephone survey.

- Includes computed weights
- Includes hundreds of variables
- Available annually (1987 to 2017)
- Very large with approximately 400,000 observations per year
- Codebook and survey available
- Many peer-reviewed papers as well as reports written based on these data
- Information about limitations and strengths included

After going to the site: http://www.cdc.gov/brfss/, download the SAS .xpt file (will take a few minutes) and you will have a "Zip" file or a .zip. Inside of the .zip file is the export file or the .XPT file. Drag and drop the .XPT file into a BRFSS directory you create on your network/computer. The transport file will be read in with the proc copy and output to the directory indicated with the libname "dataout".

```
LIBNAME TRANSPRT XPORT 'C:\YOUR PATHWAY\BRFSS\CDBRFS08.XPT';
LIBNAME DATAOUT 'C:\ YOUR PATHWAY\BRFSS\' ;
PROC COPY IN=TRANSPRT OUT=dataout;
RUN;
```

The California Health Interview Survey (CHIS) is the nation's largest state health survey with robust samples of Latinos, Asians, and American Indians.

- Includes computed weights
- Includes hundreds of variables
- Serial cross-sections every two years (2001 to 2016)
- Very large with approximately 40,000 adults per year
- Also surveys adolescents and children
- Codebook and survey available
- Many peer-reviewed papers as well as reports written based on these data
- Information about limitations and strengths included

After going to the site: http://www.chis.ucla.edu/, follow simple steps to download a SAS data set from the site.

ADDITIONAL SOURCES OF DATA FILES

There are many repositories that are being created to house data sets as well as portals that have been created to help find data. At press, here are some interesting finds:

https://www.kaggle.com/datasets

https://aws.amazon.com/public-datasets/

https://www.data.gov/education

http://www.data-planet.com/

https://www.assetmacro.com/market-data/

http://www.datasets.co/

https://nssdc.gsfc.nasa.gov/

https://www.census.gov/

SUMMARY

Individual level determinants of health are being leveraged to benefit patients by understanding disease patterns, high-risk attributes, hospital acquired conditions, and performance measures for specialized treatment approaches. Health analytics expertise are being sought to better translate medical data to actionable information and knowledge. An abundance of new data-based opportunities have made real-world data sets accessible for easy download and use in research as well as in the classroom. This paper highlights the importance of case studies in education that are adaptive to the needs of the students as they leverage SAS® and real-world data to answer relevant questions about current healthcare topics.

REFERENCES

Christensen, C. R. (1981). Teaching by the Case Method. Boston: Harvard Business School.

Carr, D. When PROC APPEND May Make More Sense Than the DATA STEP. Paper 085-2008, Proceedings of the SAS Global Forum 2008. Cary, NC: SAS Institute, Inc

Kuligowski A.T., et al. An Introduction to SAS® Arrays. Paper 6406-2016, Proceedings of the SAS Global Forum 2016. Cary, NC: SAS Institute, Inc.
https://support.sas.com/resources/papers/proceedings16/6406-2016.pdf

Waller, J. How to Use ARRAYs and DO Loops: Do I DO OVER or Do I DO i?. Proceedings of the SAS Global Forum 2010 Conference. Cary, NC: SAS Institute, Inc.
http://support.sas.com/resources/papers/proceedings10/158-2010.pdf

Pasta D.J. Learning When to Be Discrete: Continuous vs. Categorical Predictors. Paper 248–2009. Proceedings of SAS Global Forum 2009.
http://support.sas.com/resources/papers/proceedings09/248-2009.pdf

Shoemaker, JN. PROC FORMAT in Action. Paper 56-27. Proceedings of the SUGI 27 Conference. SAS Institute, Inc.
https://support.sas.com/resources/papers/proceedings/proceedings/sugi27/p056-27.pdf

Smith TC and Smith B. Using Proc Logistic and Weighting of Public Use Data in the Classroom. Paper 854-2017, Proceedings of the SAS Global Forum 2017 Conference. Cary, NC: SAS Institute, Inc. https://support.sas.com/resources/papers/proceedings17/0854-2017.pdf

Lohr SL. Using SAS® for the Design, Analysis, and Visualization of Complex Surveys. Paper 343-2012, Proceedings of the SAS Global Forum 2012 Conference. Cary, NC: SAS Institute, Inc. http://support.sas.com/resources/papers/proceedings12/343-2012.pdf

Berglund PA. Enhanced Data Analysis using SAS® ODS Graphics and Statistical Graphics. Paper 343-2012, Proceedings of the SAS Global Forum 2012 Conference. Cary, NC: SAS Institute, Inc.

Lewis T. Considerations and Techniques for Analyzing Domains of Complex Survey Data. Paper 449-2013, Proceedings of the SAS Global Forum 2013 Conference. Cary, NC: SAS Institute, Inc. http://support.sas.com/resources/papers/proceedings13/449-2013.pdf

Cassell D. Wait Wait, Don't Tell Me… You're Using the Wrong Proc! Paper 193-31, SUGI 31. https://support.sas.com/resources/papers/proceedings/proceedings/sugi31/193-31.pdf

Rheta E. Lanehart et al. Propensity Score Analysis and Assessment of Propensity Score Approaches using SAS® Procedures. Paper 314-2012, Proceedings of SAS Global Forum 2012. https://pdfs.semanticscholar.org/c10a/cf956fd8fd91debda40aa16c18f707303ec7.pdf

Cepeda, et al. Comparison of Logistic Regression versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders. Am J Epidemiology 2003.

Austin. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research, 46:399–424, 2011.

ACKNOWLEDGMENTS

ABOUT THE AUTHORS AND CONTACT INFORMATION

Dr. Tyler Smith is professor of biostatistics, epidemiology, public health and health informatics; and program director for the Health Analytics master's degree. Dr. Smith received a BS in mathematics/statistics from California State University, Chico; MS in statistics from the University of Kentucky; and PhD in epidemiology from the University of California, San Diego. With >20 years of experience in health research leading large longitudinal studies, infant health registries, and medical health outcomes research, he has ~150 peer-reviewed publications in scientific journals, ~300 scientific presentations and has been PI/COI on grants totaling >$20,000,000. Currently Dr. Smith has served the SAS community through his efforts as Statistics Chair and Content Area Lead for SAS Global Forum for many years; 2015 SAS Global Forum Conference Chair; Academic Program Chair for Western User's of SAS Software, and as

part of the Executive Board for the San Diego SAS User's Group, SAS Global Forum, and Western User's of SAS Software.

Tyler C Smith, MS, PhD
Professor and Chair
Program Director MS Health and Life Science Analytics
Director Health Science Research Center
Department of Community Health
School of Health and Human Services
National University
San Diego, CA 92123
tsmith@nu.edu

Dr. Besa Smith has worked in government, academic, and private industries and has served as a senior epidemiologist, senior biostatistician, and head of analytics for a 35-40 member multi-disciplinary research team. She has taught epidemiology and biostatistics courses to undergraduate, graduate, and medical students and founded Analydata, a statistical consulting group. She is currently Global Head of Biostatistics and Medical Writing, ICON Commercialisation and Outcomes, ICON, plc. Dr. Smith has a BS in biology; MPH in biometry, and PhD in epidemiology. With nearly 20 years leveraging health analytics in longitudinal studies and medical health outcomes research, she has ~80 peer-reviewed publications in scientific journals and >100 scientific presentations. Dr. Smith has served the SAS community as Statistics Chair and Content Area Lead for SAS Global Forum for many years; Academic Program Chair for Western User's of SAS Software, and as part of the Executive Board for the San Diego SAS User's Group, and Western User's of SAS Software.

Besa Smith, MPH, PhD
Global Lead
Biostatistics and Medical Writing
Medical Affairs Statistical Analysis
Real World Evidence Late Phase Research
ICON Commercialisation & Outcomes
Besa.Smith@iconplc.com