SAS Analytics Guide:

How to perform binary logistic regression

Data version: Controlled Tier *All of Us* Curated Data Repository (CDR) v7 (2022Q4R11)
Analysis tool: SAS Studio
Authors:
Emily Goldmann, Ph.D., MPH (✉)
New York University School of Global Public Health

Stephanie Cook, Dr.Ph., MPH (✉)
New York University School of Global Public Health

Please note: This guide aims to demonstrate the utilization of **PROC FREQ** and **PROC LOGISTIC** procedures using SAS Studio. However, it is important to note that this guide is not comprehensive and does not encompass all facets of the scientific process which researchers are required to undertake nor does it assume that this is the only way to correctly perform these statistical procedures. Specifically, it does not delve into data cleaning and verification, assumption validation, model diagnostics, potential follow-up analyses, or any other possible approaches for performing these frequency and binary logistic regression procedures.

Table of Contents

Introduction

This guide includes examples of statistical analysis processes which were used to assess concordance between self-reported lifetime depression diagnosis and depressive disorder diagnoses documented in available electronic health records (EHR) using survey and EHR data from the *All of Us* dataset.

More specifically, these examples demonstrate the prevalence and demographic correlates of ***not*** self-reporting a lifetime depression diagnosis among adult (18 years or older) respondents with evidence of depressive disorder in EHR.

The study described in this guide is descriptive in nature and provides overall prevalence of not self-reporting depression among those with EHR depressive disorder and demographic patterns within this prevalence. To describe these categorical variables, frequencies and percentages are reported.

To evaluate bivariable associations between self-reporting depression and demographic factors, a Pearson's chi-square test is generally performed. Finally, one can employ a binary logistic regression model including all demographic variables as independent variables to identify those

factors that are *independently* associated with the outcome, i.e., not self-reporting a lifetime depression diagnosis.

This type of regression is commonly used for binary (yes/no) dependent variables and yields odds ratios (OR) as a relative measure of association between the dependent variable and each independent variable, controlling for all other independent variables in the model. 95% confidence intervals (CI) also accompany each OR as an indication of estimated precision.

To begin our discussion, it is important to outline some definitions of terms frequent used throughout this guide:

- ***Outcome or independent variable***: A variable of interest for which we want to understand overall prevalence (or proportion) in a population and might vary across demographic subgroups (dependent variables).
- ***Dependent variable***: A variable of interest for which we want to understand its association with the outcome variable. In descriptive studies, these variables are often referred to as correlates.

Description of the dataset

For this example, we identified two cohorts for analysis:

1. **Case group:** Respondents who had an electronic health record (EHR) depression code but did not self-report depression
    - Adults (Participants aged 18 years or older)
    - Did not report that they had a lifetime depression diagnosis in the Personal and Family Health History survey
    - Had at least one EHR diagnostic code for depressive disorder (SNOMED code: 35489007, depressive disorder)
2. **Control group:** Respondents who had an EHR depression code and self-reported depression
    - Adults (Participants aged 18 years or older)
    - Reported that they had a lifetime depression diagnosis in the Personal and Family Health History survey
    - Had at least one EHR diagnostic code for depressive disorder (SNOMED code: 35489007, depressive disorder)

Demographic correlates / dependent variables included:

- *Date_of_birth*, to calculate age in years at the time of data analysis (1=18-44, 2=45-64, 3=65-84, 4=85 or older)
    - **Recoded variable name: age_cat**
- *Race* and *Ethnicity*, to generate a variable that combines race and ethnicity (1=Non-Hispanic white, 2=Non-Hispanic Black, 3=Hispanic any race, 4=Non-Hispanic Asian, Native Hawaiian or other Pacific Islander, 5=Non-Hispanic multiple or other race)
    - **Recoded variable name: race_eth**

- *Gender* (1=Female, 2=Male, 3=Other)
    - **Recoded variable name: gender_new**
- *Sexual Orientation* (1=Straight, 2=Gay or Lesbian, 3=Bisexual, 4=Other)
    - **Recoded variable name: sex_orient**
- *Highest Education Level* (1=Less than high school, 2=High school graduate/GED, 3=Some college, 4=College graduate, 5=Advanced degree)
    - **Recoded variable name: edu_cat**

Other variables:

- *Person_id* as a unique participant identifier, used to merge and deduplicate datasets

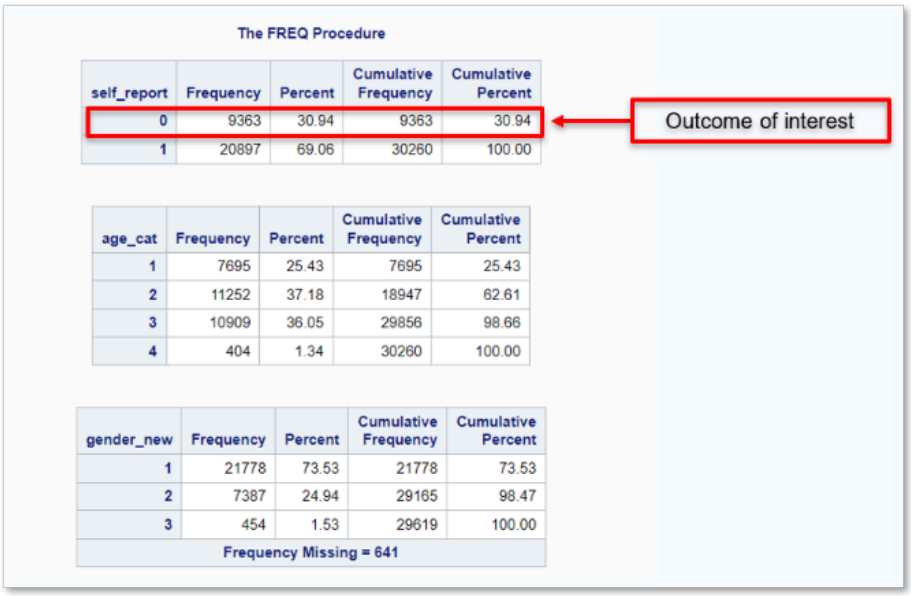
Statistical analysis procedures

Step 1: Describe all variables using frequencies and percentages among adults with an electronic health record (EHR) depression diagnosis (n=30,260).

Example code:

```
proc freq data=mydata.depress_final2;
tables self_report age_cat gender_new race_eth sex_orient edu_cat;
run; * 30.94% have EHR Depression Dx but did not self-report depression Dx;
```

Example results:



The FREQ Procedure

self_report	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	9363	30.94	9363	30.94
1	20897	69.06	30260	100.00

Outcome of interest

age_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7695	25.43	7695	25.43
2	11252	37.18	18947	62.61
3	10909	36.05	29856	98.66
4	404	1.34	30260	100.00

gender_new	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	21778	73.53	21778	73.53
2	7387	24.94	29165	98.47
3	454	1.53	29619	100.00

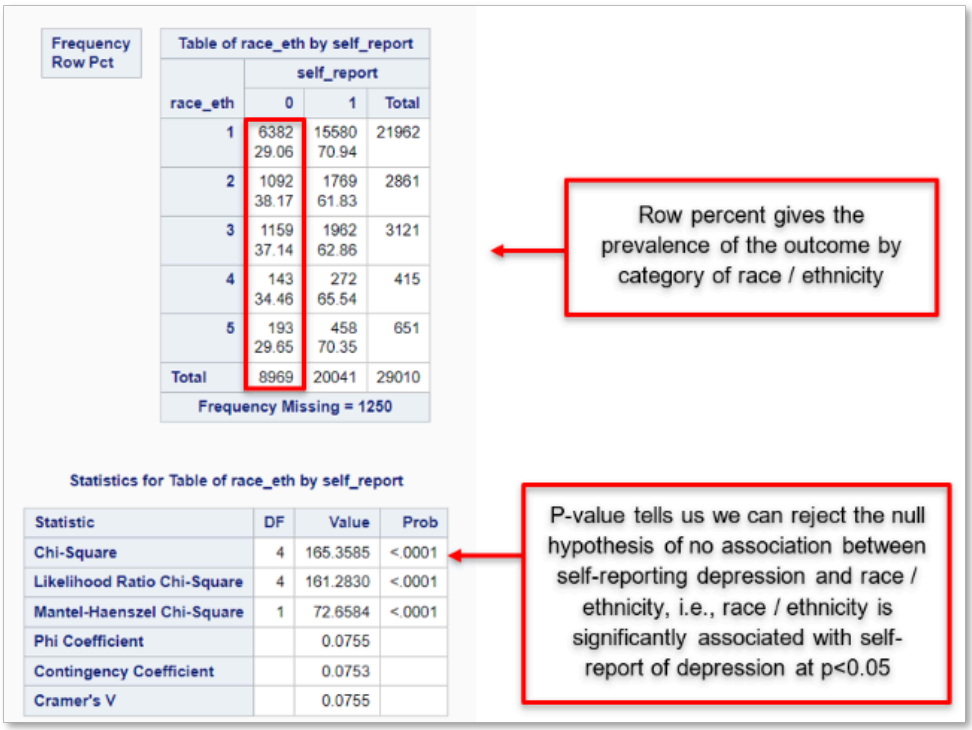
Frequency Missing = 641

Step 2: Examine associations between the outcome (not self-reporting depression diagnosis) and each demographic variable among adults with an EHR depression diagnosis (n=30,260) using cross tabulation and Pearson's Chi-square test.

Example code:

```
proc freq data=mydata.depress_final2;
tables (age_cat gender_new race_eth sex_orient edu_cat)*self_report / chisq nopercent nocol;
run;
```

Example results:



Frequency
Row Pct

Table of race_eth by self_report

race_eth	self_report 0	1	Total
1	6382 29.06	15580 70.94	21962
2	1092 38.17	1769 61.83	2861
3	1159 37.14	1962 62.86	3121
4	143 34.46	272 65.54	415
5	193 29.65	458 70.35	651
Total	8969	20041	29010

Frequency Missing = 1250

Row percent gives the prevalence of the outcome by category of race / ethnicity

Statistics for Table of race_eth by self_report

Statistic	DF	Value	Prob
Chi-Square	4	165.3585	<.0001
Likelihood Ratio Chi-Square	4	161.2830	<.0001
Mantel-Haenszel Chi-Square	1	72.6584	<.0001
Phi Coefficient		0.0755	
Contingency Coefficient		0.0753	
Cramer's V		0.0755	

P-value tells us we can reject the null hypothesis of no association between self-reporting depression and race / ethnicity, i.e., race / ethnicity is significantly associated with self-report of depression at p<0.05

These results suggest that among adults 18 years or older with EHR depression diagnosis, non-Hispanic Black (38.17%), Hispanic any race (37.14%), and non-Hispanic Asian (34.46) respondents had a higher prevalence of not self-reporting depression diagnosis compared to non-Hispanic white respondents (29.06%, p<0.0001).

Another example evaluates the association between the outcome and the *intersection* of two demographic factors, e.g., association between the outcome and race/ethnicity by gender.

Example code:

```
proc freq data=mydata.depress_final2;
tables gender_new*race_eth*self_report / chisq nopercent nocol;
run;
```

Example results:

For women only (**gender_new=1**)

Frequency
Row Pct

Table 1 of race_eth by self_report
Controlling for gender_new=1

race_eth	self_report 0	1	Total
1	4506 28.50	11303 71.50	15809
2	860 37.95	1406 62.05	2266
3	907 37.57	1507 62.43	2414
4	102 36.17	180 63.83	282
5	144 29.94	337 70.06	481
Total	6519	14733	21252

Frequency Missing = 526

Statistics for Table 1 of race_eth by self_report
Controlling for gender_new=1

Statistic	DF	Value	Prob
Chi-Square	4	149.6463	<.0001
Likelihood Ratio Chi-Square	4	145.9308	<.0001
Mantel-Haenszel Chi-Square	1	75.2131	<.0001
Phi Coefficient		0.0839	
Contingency Coefficient		0.0836	

For men only (**gender_new=2**)

Frequency
Row Pct

Table 2 of race_eth by self_report
Controlling for gender_new=2

race_eth	self_report 0	1	Total
1	1814 31.66	3915 68.34	5729
2	219 39.60	334 60.40	553
3	239 36.88	409 63.12	648
4	39 33.62	77 66.38	116
5	46 31.94	98 68.06	144
Total	2357	4833	7190

Frequency Missing = 197

Statistics for Table 2 of race_eth by self_report
Controlling for gender_new=2

Statistic	DF	Value	Prob
Chi-Square	4	19.9542	0.0005
Likelihood Ratio Chi-Square	4	19.5197	0.0006
Mantel-Haenszel Chi-Square	1	6.6218	0.0101
Phi Coefficient		0.0527	
Contingency Coefficient		0.0526	
Cramer's V		0.0527	

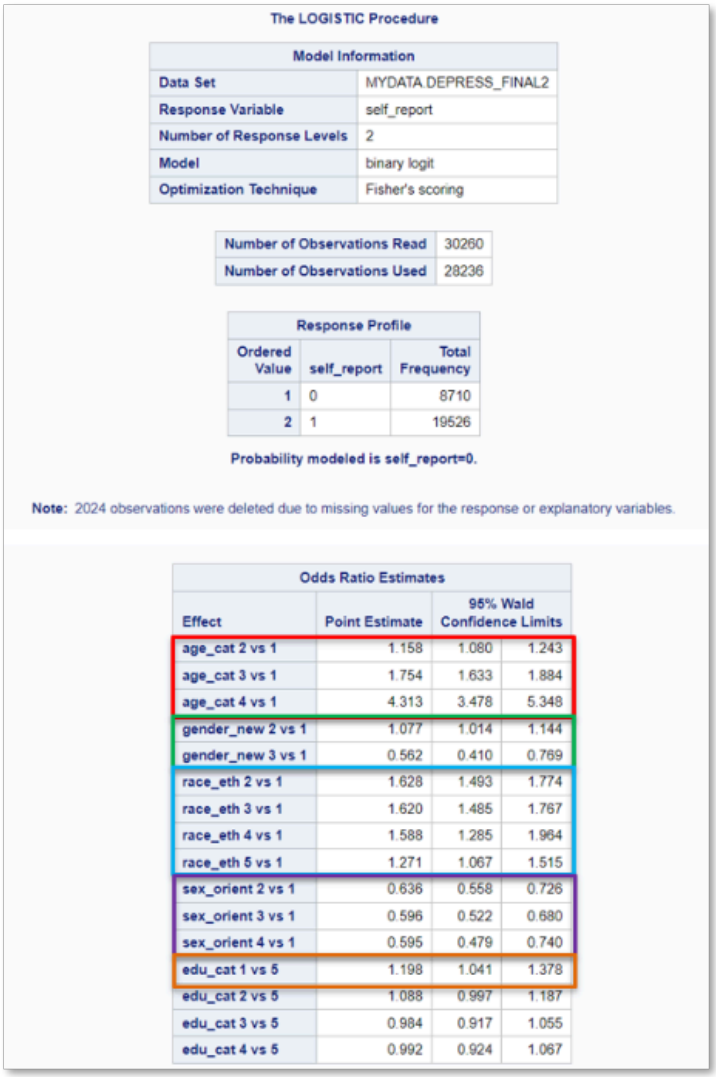
This yielded a similar pattern in the prevalence of the outcome by race/ethnicity among women and men.

Step 3: Identify demographic variables independently associated with not self-reporting depression among adults with an EHR depression diagnosis.

Example code:

```
proc logistic data=mydata.depress_final2;
class age_cat (ref="1") gender_new (ref="1") race_eth (ref="1") sex_orient (ref="1") edu_cat
(ref="5");
model self_report = age_cat gender_new race_eth sex_orient edu_cat;
run;
```

Example results:



The LOGISTIC Procedure

Model Information	
Data Set	MYDATA.DEPRESS_FINAL2
Response Variable	self_report
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	30260
Number of Observations Used	28236

Response Profile

Ordered Value	self_report	Total Frequency
1	0	8710
2	1	19526

Probability modeled is self_report=0.

Note: 2024 observations were deleted due to missing values for the response or explanatory variables.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age_cat 2 vs 1	1.158	1.080	1.243
age_cat 3 vs 1	1.754	1.633	1.884
age_cat 4 vs 1	4.313	3.478	5.348
gender_new 2 vs 1	1.077	1.014	1.144
gender_new 3 vs 1	0.562	0.410	0.769
race_eth 2 vs 1	1.628	1.493	1.774
race_eth 3 vs 1	1.620	1.485	1.767
race_eth 4 vs 1	1.588	1.285	1.964
race_eth 5 vs 1	1.271	1.067	1.515
sex_orient 2 vs 1	0.636	0.558	0.726
sex_orient 3 vs 1	0.596	0.522	0.680
sex_orient 4 vs 1	0.595	0.479	0.740
edu_cat 1 vs 5	1.198	1.041	1.378
edu_cat 2 vs 5	1.088	0.997	1.187
edu_cat 3 vs 5	0.984	0.917	1.055
edu_cat 4 vs 5	0.992	0.924	1.067

All of the odds ratios (OR) highlighted above show an estimate of the odds of the outcome (not self-reporting depression diagnosis) in that group that is significantly different from the odds of the outcome in the reference group (OR>1 if the odds are higher, OR<1 if the odds are lower). 95% confidence intervals (CI) that do not include the value of 1.000 indicate statistically significant associations at $p<0.05$.

For example, we see that compared to non-Hispanic white respondents with an electronic health record (EHR) depression diagnosis (race_eth=1, the reference group), respondents in most other race/ethnicity groups had significantly higher odds of not self-reporting depression diagnosis (e.g., compared to non-Hispanic white respondents with EHR depression diagnosis, non-Hispanic Black (race_eth=2) respondents with EHR depression diagnosis had approximately 63% higher odds of not self-reporting depression diagnosis; OR=1.628, 95% CI: 1.493-1.774).

This suggests that non-Hispanic Black respondents with an EHR depression diagnosis are more likely than non-Hispanic white respondents with an EHR depression diagnosis to not self-report lifetime depression diagnosis, despite EHR documentation of depression diagnosis.

Additional resources

For additional information about using SAS Studio in the Researcher Workbench, explore the following articles: Exploring All of Us data using SAS Studio and How to run SAS in the Researcher Workbench.

Updated April 9, 2024