`predict` and `margins`

Andy Grogan-Kaylor

23 Sep 2020 14:35:15

Background

Odds ratios, or coefficients showing the association of the independent variables with the log odds, represent the most immediate output of a logistic regression. However, for a variety of reasons, it may make sense to not only report odds ratios, but also to investigate predicted probabilities.

The Data

The data are an extract of the *National Survey of Children's Health, 2018.* The data contain information on children's current depression status, their exposure to various *Adverse Childhood Experiences* (ACEs) and their sex and race.

```
. clear all

. cd "/Users/agrogan/Desktop/newstuff/categorical/predict-and-margins"
/Users/agrogan/Desktop/newstuff/categorical/predict-and-margins

. use "NSCH_ACES.dta", clear

. describe
Contains data from NSCH_ACES.dta
  obs:        30,530
  vars:           13                          23 Sep 2020 13:49
  size:       488,480
─────────────────────────────────────────────────────────────────────────────
              storage   display    value
variable name   type    format     label       variable label
─────────────────────────────────────────────────────────────────────────────
sc_sex          byte    %30.0g     sc_sex_lab
                                               Sex of Selected Child
sc_race_r       byte    %48.0g     sc_race_r_lab
                                               Race of Selected Child, Detailed
sc_racer        byte    %31.0g     sc_racer_lab
                                               Race of Selected Child, Recode
depress         float   %9.0g
ace1R           byte    %9.0g                  RECODE of ace1 (Hard to Cover Basics Like Food
                                                  or Housing)
ace3R           byte    %9.0g                  RECODE of ace3 (Child Experienced - Parent or
                                                  Guardian Divorced)
ace4R           byte    %9.0g                  RECODE of ace4 (Child Experienced - Parent or
                                                  Guardian Died)
ace5R           byte    %9.0g                  RECODE of ace5 (Child Experienced - Parent or
                                                  Guardian Time in Jail)
ace6R           byte    %9.0g                  RECODE of ace6 (Child Experienced - Adults
                                                  Slap, Hit, Kick, Punch Others)
ace7R           byte    %9.0g                  RECODE of ace7 (Child Experienced - Victim of
                                                  Violence)
ace8R           byte    %9.0g                  RECODE of ace8 (Child Experienced - Lived with
```

```
                                                      Mentally Ill)
       ace9R           byte   %9.0g          RECODE of ace9 (Child Experienced - Lived with
                                                Person with Alcohol/Drug Problem)
       ace10R          byte   %9.0g          RECODE of ace10 (Child Experienced - Treated
                                                Unfairly Because of Race)
       ─────────────────────────────────────────────────────────────────────────────────────

       Sorted by:
```

Logistic Regression

We estimate a logistic regression using `,or` to ask for *odds ratios*.

```
. logit depress ace1R ace3R ace4R ace5R ace6R ace7R ace8R ace9R ace10R i.sc_race_r i.sc_sex,
>  or

Iteration 0:   log likelihood = -4742.8248
Iteration 1:   log likelihood = -4256.2811
Iteration 2:   log likelihood = -4180.3512
Iteration 3:   log likelihood = -4179.6624
Iteration 4:   log likelihood =  -4179.661
Iteration 5:   log likelihood =  -4179.661

Logistic regression                             Number of obs   =     28,926
                                                LR chi2(16)     =    1126.33
                                                Prob > chi2     =     0.0000
Log likelihood =  -4179.661                     Pseudo R2       =     0.1187
```

depress	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf. Interval]	
ace1R	1.611495	.1106592	6.95	0.000	1.408569	1.843656
ace3R	1.627997	.1212124	6.55	0.000	1.406946	1.883778
ace4R	1.78926	.2211975	4.71	0.000	1.404247	2.279835
ace5R	1.048195	.1097253	0.45	0.653	.8537646	1.286904
ace6R	1.136375	.1253184	1.16	0.246	.9154867	1.41056
ace7R	1.9394	.2129262	6.03	0.000	1.563919	2.40503
ace8R	3.72644	.2966351	16.53	0.000	3.188131	4.355642
ace9R	1.410608	.1304044	3.72	0.000	1.176837	1.690815
ace10R	1.535313	.2020991	3.26	0.001	1.186178	1.98721
sc_race_r						
Black or African Ameri..	.827289	.1074881	-1.46	0.144	.6413015	1.067216
American Indian or Ala..	.6279865	.2045843	-1.43	0.153	.3316252	1.189195
Asian alone	.7145884	.1439746	-1.67	0.095	.4814561	1.060609
Native Hawaiian and Ot..	.2644952	.2745993	-1.28	0.200	.03457	2.023654
Some Other Race alone	.6443303	.1434801	-1.97	0.048	.4164493	.9969076
Two or More Races	.616392	.0814335	-3.66	0.000	.4757754	.7985681
sc_sex						
Female	1.468479	.0935905	6.03	0.000	1.296039	1.663862
_cons	.0150195	.0010428	-60.47	0.000	.0131087	.0172089

```
Note: _cons estimates baseline odds.
```

Predicted Probabilities

Predicted probabilities are *each participant's individual predicted probability* of experiencing depression based upon the independent variables included in the model. We often denote such predicted probabilities with $\hat{y}$

```
. predict yhat
(option pr assumed; Pr(depress))
(1,604 missing values generated)
```

`yhat` is a variable in the data, just like any other variable, and we can tabulate and graph it.

```
. tabulate sc_race_r, summarize(yhat)
```

```
     Race of
   Selected
     Child,          Summary of Pr(depress)
   Detailed         Mean   Std. Dev.        Freq.

  White alo     .04050317   .05252093       22,418
  Black or      .04153355   .05702566        1,878
  American      .05128205   .08430179          234
  Asian alo     .01963636   .01999562        1,375
  Native Ha     .01369863   .02790487           73
  Some Othe     .03038309   .04326265          757
  Two or Mo     .03286171   .05022293        2,191

     Total      .03875406   .05190828       28,926
```

```
. graph bar yhat, ///
> over(sc_race_r, label(angle(forty_five))) ///
> title("Predicted Probability of Depression") ///
> scheme(michigan)

. graph export mybar.png, width(500) replace
(file mybar.png written in PNG format)
```
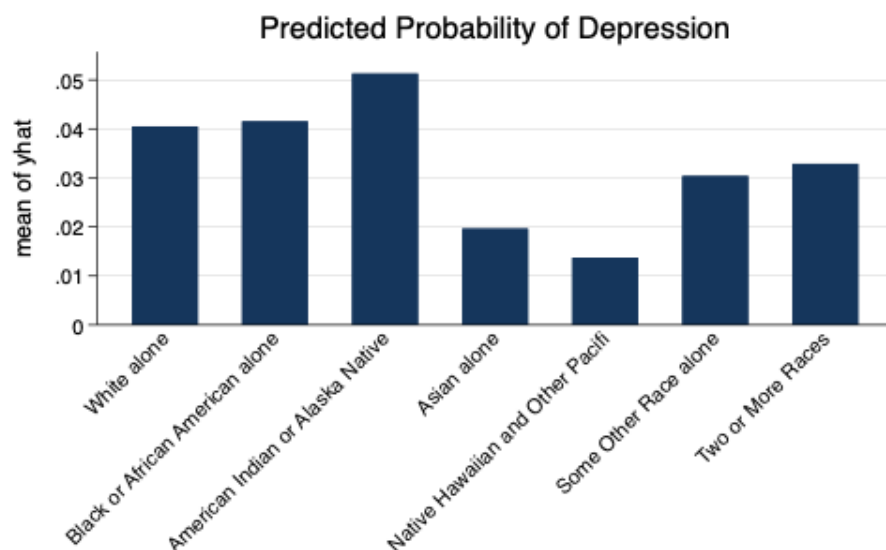

Figure 1: Bar Graph of Predicted Probabilities

Predicted Margins (Over A Variable of Interest)

In their simplest form, predictive margins are *average predicted probabilities* were everyone in the sample were treated as if they were of a particular race.

```
. margins sc_race_r // predictive margins

Predictive margins                              Number of obs     =     28,926
Model VCE    : OIM

Expression   : Pr(depress), predict()

                             Delta-method
                   Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

3

sc_race_r						
White alone	.0415443	.0013092	31.73	0.000	.0389785	.0441102
Black or African Ameri..	.0350812	.0038971	9.00	0.000	.0274431	.0427194
American Indian or Ala..	.0273233	.0080773	3.38	0.001	.0114921	.0431546
Asian alone	.0307367	.0055275	5.56	0.000	.019903	.0415704
Native Hawaiian and Ot..	.0121547	.0120465	1.01	0.313	-.0114559	.0357653
Some Other Race alone	.0279727	.0056061	4.99	0.000	.016985	.0389605
Two or More Races	.0268611	.0031053	8.65	0.000	.0207748	.0329474

We could also evaluate `margins` holding other variables at their *mean* values using the `atmeans` option. You can also read about obtaining `margins` for various combinations of the independent variables by typing `help margins` at the Stata prompt.

The essential graphing command is `marginsplot`, which will usually produce a perfectly useable graph. The other graphing options are added for clarification and aesthetic purposes.

```
. marginsplot, ///
> title("Predicted Probability of Depression") ///
> ylabel(, labsize(small) angle(horizontal)) ///
> xlabel(, angle(forty_five)) ///
> scheme(michigan)

  Variables that uniquely identify margins: sc_race_r

. graph export mymargins.png, width(500) replace
(file mymargins.png written in PNG format)
```
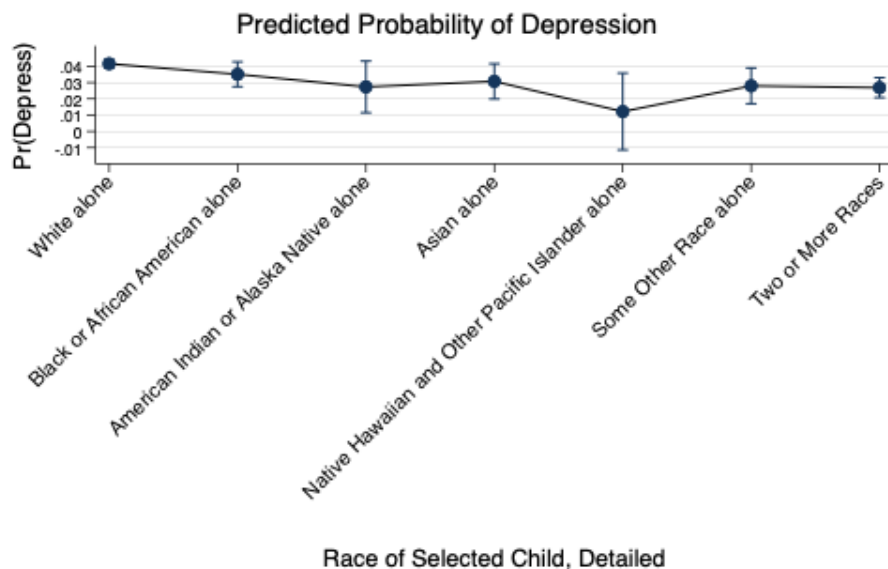

Figure 2: Margins Plot of Predicted Probabilities