Soc 73994, Homework #7
Models for Binary Outcomes II & III

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised October 25, 2024

All answers should be submitted through Canvas. Be sure your response includes your name, the date, and a clear title, e.g. Homework # 7. If there is a huge amount of output for any analyses you run yourself, you may want to be selective in what you copy and paste into your assignment (but make sure you include enough so it is clear what commands you executed, e.g. you might show all the commands but only parts of the output).

This assignment focuses primarily on comparing logit and probit coefficients across nested models. First you will interpret the results from a prepared problem, and you will then do a similar anlysis using a data set of your choice. Ideally you will use your own data for the latter but you can use something else. Note that, if you normally use the svy: prefix, it may not work with the commands you are asked to use so you may have to drop the prefix for this assignment. (If dropping svy: makes you nervous, you could try running your logistic regressions both with and without svy: and see if it makes much difference.)

1. Prepared problem. I've already told you most of the answers to this question! But I want you do it anyway so you understand how it works and can explain things in your own words.

We will once again use the American National Election Study, 2008 (ANES 2008). We will NOT weight the data because the khb command does not support weighting, but luckily the logistic regression coefficients and the marginal effects are very similar whether you use svy: or not. Run the following commands:

```
clear all
use https://www3.nd.edu/~rwilliam/statafiles/anes_codeddata, clear
* estout, fre, khb, & spost13_ado need to be installed

********** Prepare the data first! **********
* Clean up income variables & some labels
mvdecode V083309A, mv(-4 = .a \ 98 = .b)
label variable income "Estimated Family Income in $1000s (recode of V083309A)"
label variable pres2008 "Who did R vote for in 2008? (0 = McCain, 1 = Obama)"
* pres2008 is a recode of the variable president, with the 25
* who voted for other candidates recoded to missing.

* We aren't going to use weights in this problem, but we'll
* leave the weights in in case you want to double-check whether the
* logistic regression coefficients and the marginal effects
* are affected much by whether you use or do not use weights.
svyset CASEID [pweight = V080102A]

* Dichotomize race to make analysis simpler. The nonwhite groups all voted
```

```
* for Obama by wide margins so combining them seems reasonable.
recode race (1 = 1 "White") (2 3 4 = 0 "NonWhite") (else = .), gen(white) label (white)
label variable white "race recoded to 0 = NonWhite, 1 = White"

* For convenience, we will keep ONLY the variables and cases we will be using.
* DON'T drop the cases if you want to use svy though, as standard
* errors will be off.
keep CASEID V080102A pres2008 white age income bush feminist V083309A
keep if !missing(CASEID, V080102A, pres2008, white, age, income, bush,
feminist, V083309A)

* Descriptive information on the variables used. Descriptive stats are off a bit
* because weighting is not being used. e.g. Obama's margin of victory
* was NOT as large as the unweighted numbers imply.
codebook, compact
fre pres2008 white

********** Now do the actual analysis! **********
capture estimates drop m1 m2
* We won't use weights because the khb command does not support them.
* Luckily the logistic regression coefficients and margins results
* are similar whether you use svy: or do not use it.

* 1a. First, compare coefficients across nested models:
logit pres2008 i.white, nolog
est store m1
logit pres2008 i.white age income bush feminist, nolog
est store m2
esttab m1 m2, t scalars(r2_p chi2 df_m p bic) sfmt(%9.4f) obslast nobaselevels

* 1b. Now look at the y-standardized coefficients, given in the bStdY column,
* and see if they lead to the same or different conclusions about
* how the effect of white changes as more variables are added to the model.
est restore m1
listcoef, std
est restore m2
listcoef, std

* 1c. Now look at the khb results, and see if they lead to the
* same or different conclusions as the original coefficients did about
* how the effect of white changes as more variables are added to the model.
khb logit pres2008 i.white || age income bush feminist, nolog

* 1e. Now look at the marginal effects. In this case it doesn't matter much
* whether you the margins command, the MDL approach, or khb.
estimates restore m1
margins, dydx(white)
estimates restore m2
margins, dydx(white)
khb logit pres2008 i.white || age income bush feminist, nolog ape
```

Discuss the following:

a. How does the coefficient for the variable in the baseline model, **white**, change as more variables are added. In particular, does the coefficient for white increase or decrease as variables are added? Just looking at the coefficients for **white** in the two models, does the change seem large or small to you?

b. How do the y-standardized coefficient coefficients for **white** change as more variables are added? Just looking at the y-standardized coefficients for **white** in the two models, does the change between them seem large or small to you?

c. What does the khb command tell you about the changes in the effect of **white** across models? Does khb make the change seem large or small to you? Does the indirect effect of **white** seem fairly small or fairly large to you?

d. Briefly explain any discrepancies between the conclusions you reached in (a) about the magnitude of the change in the white coefficient versus what you concluded in (b) and (c). Specifically, discuss how and why y* gets rescaled in the 2nd model. Then discuss how and why your comparison of coefficients across models may be affected by the rescaling. Remember that listcoef gives you the standard deviation of y*, so you will want to mention that in your discussion.

e. Finally, what does the last khb command tell you about how the marginal effect of **white** changes across models? Does the change seem fairly large to you or does it seem fairly trivial? Would you have expected such a large (or small) change based on the original logistic coefficients? (Note: The Mize/Doan/Long approach yields very similar results, and also shows that the change in marginal effects between models is statistically significant.)

For the remaining problems use a data set of your choice.

2. Run a series of at least two nested logistic regression models.

a. The first model should be very simple, maybe even only one variable (e.g. race, gender). The 2nd model should add a few variables, and any additional models should add even more. There should be some sort of logical sequencing to the order in which you add variables. Ideally the added variables will be statistically significant and noticeably increase pseudo R^2. Further these variables should be expected to account for at least some of the effects of the variables in your baseline model, e.g. the direct effect of race should decline as education is controlled for. Make sure the sample used is the same for each model. If you are not sure how to do this see the first few pages of the Missing Data Part 1 handout.

b. Use something like esttab, outreg2, or estimates table to display your models in side by side columns. Discuss how the coefficients in your first model go up or down as more variables are added. Just looking at the coefficients, do the changes seem large or small to you?

3. After each model, run the command listcoef, std help. Then discuss

a. How the y-standardized coefficients change as more variables are added.

b. Briefly explain why y* gets rescaled with each equation. Then discuss how and why your comparison of coefficients across models may be affected by the rescaling. Remember that listcoef gives you the standard deviation of y*, so you will want to mention that in your discussion.

4. Now use the khb command.

a. Choose a variable (or variables) from your first model. What does the khb commnd tell you about how the coefficients in your first model change as more variables are added? (You can also use khb to examine how the marginal effects change between models if you want.) What does the khb commnd tell you about the direct and indirect effects of that variable?

b. In your case, do you think the khb command is helping you to make (at least slightly) more correct statements than you would have made had you only looked at how the coefficients changed across models? (Don't feel too bad if the results don't seem too dramatic! That is often the case. But if you do have a reasonable example where KHB seems to make a major difference in how you interpret the findings, I would like to see it.)

5. [Optional] Discuss anything else you think is useful from your analyses, e.g. do the x-standardized coefficients from your final model tell you much?