Comparing Logit & Probit Coefficients Between Nested Models (Old Version)

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/

Last revised March 24, 2019

NOTE: Long and Freese's `spost9` programs are used in this handout; specifically, the `listcoef` command, which is part of `spost9`, is used. Use the `findit` command to locate and install `spost9`. See Long and Freese's book, Regression Models for Categorical Dependent Variables Using Stata, Second Edition, for more information. Long's 1997 Regression Models for Categorical and Limited Dependent Variables provides a brief substantive discussion on pp. 69-71.

Overview. Social Scientists are often interested in seeing how the effects of variables differ between models. For example, a researcher might want to know whether the estimated effect of race on some outcome declines once education is controlled for. In OLS regression with continuous dependent variables, such issues are often addressed by estimating and comparing sequences of nested models. Unfortunately, these same approaches can be highly problematic when binary and ordinal dependent variables are analyzed via probit or logistic regression. Naïve comparisons of coefficients between models can indicate differences where none exist, hide differences that do exist, and even show differences in the opposite direction of what actually exists. This handout explains the problems and discusses the strengths and weaknesses of various proposed solutions, including Y-standardization (Winship & Mare, 1984).

Example. An example will illustrate the problems. I have constructed a data set such that x1 and x2 are uncorrelated with each other. Both have strong effects on y. ybinary is a dichotomized version of y, where y values above 0 are recoded to 1 and values of 0 and below are recoded to 0. Let's compare the results of OLS regression and logistic regression.

```
. use https://www3.nd.edu/~rwilliam/statafiles/standardized.dta
. corr, means

(obs=500)

    Variable |         Mean    Std. Dev.          Min          Max
-------------+--------------------------------------------------
           y |     5.51e-07     3.000001    -8.508021     7.981196
     ybinary |         .488     .5003566            0            1
          x1 |    -2.19e-08            2     -6.32646     6.401608
          x2 |     3.57e-08            3    -10.56658     9.646875


             |        y  ybinary       x1       x2
-------------+------------------------------------
           y |   1.0000
     ybinary |   0.7923   1.0000
          x1 |   0.6667   0.5248   1.0000
          x2 |   0.6667   0.5225   0.0000   1.0000
```

```
. qui reg y x1
. listcoef, std

regress (N=500): Unstandardized and Standardized Estimates

 Observed SD: 3.0000014
 SD of Error: 2.2383128

------------------------------------------------------------------------------
           y |      b         t     P>|t|    bStdX    bStdY   bStdXY      SDofX
-------------+----------------------------------------------------------------
          x1 |   1.00000   19.960   0.000   2.0000   0.3333   0.6667      2.0000
------------------------------------------------------------------------------

. qui reg y x2
. listcoef, std

regress (N=500): Unstandardized and Standardized Estimates

 Observed SD: 3.0000014
 SD of Error: 2.2383131

------------------------------------------------------------------------------
           y |      b         t     P>|t|    bStdX    bStdY   bStdXY      SDofX
-------------+----------------------------------------------------------------
          x2 |   0.66667   19.960   0.000   2.0000   0.2222   0.6667      3.0000
------------------------------------------------------------------------------

. qui reg y x1 x2
. listcoef, std

regress (N=500): Unstandardized and Standardized Estimates

 Observed SD: 3.0000014
 SD of Error: 1.0020108

------------------------------------------------------------------------------
           y |      b         t     P>|t|    bStdX    bStdY   bStdXY      SDofX
-------------+----------------------------------------------------------------
          x1 |   1.00000   44.587   0.000   2.0000   0.3333   0.6667      2.0000
          x2 |   0.66667   44.587   0.000   2.0000   0.2222   0.6667      3.0000
------------------------------------------------------------------------------
```

As we see, in an OLS regression, when x1 and x2 are uncorrelated with each other, their metric and standardized effects are the same in the bivariate regressions as they are when y is regressed on both x's simultaneously. (Basically, this is the special case of omitted variable bias: when the x's are uncorrelated with each other, leaving one x out does not bias the estimated effect of the other.) Further, the Standard Deviation of Y is the same across models; but as more variables are added to the model the residual variance goes down.

Important! Put another way, in OLS regression with an observed variable Y, the variance of Y is a fixed quantity; it will neither increase nor decrease as variables are added to the model. What will change is that, as variables are added to the model, the explained variance will increase and the residual variance will decrease by equal amounts.

Compare this now to the results of a logistic regression:

```
. quietly logit ybinary x1
```

```
. listcoef, std

logit (N=500): Unstandardized and Standardized Estimates

 Observed SD: .50035659
   Latent SD: 2.3395663

  Odds of: 1 vs 0

------------------------------------------------------------------------------
     ybinary |      b         z     P>|z|    bStdX    bStdY   bStdXY      SDofX
-------------+----------------------------------------------------------------
          x1 |   0.73887   10.127   0.000   1.4777   0.3158   0.6316     2.0000
------------------------------------------------------------------------------

. quietly logit  ybinary x2
. listcoef, std

logit (N=500): Unstandardized and Standardized Estimates

 Observed SD: .50035659
   Latent SD: 2.3321875

  Odds of: 1 vs 0

------------------------------------------------------------------------------
     ybinary |      b         z     P>|z|    bStdX    bStdY   bStdXY      SDofX
-------------+----------------------------------------------------------------
          x2 |   0.48868   10.134   0.000   1.4660   0.2095   0.6286     3.0000
------------------------------------------------------------------------------

. quietly logit  ybinary x1 x2
. listcoef, std

logit (N=500): Unstandardized and Standardized Estimates

 Observed SD: .50035659
   Latent SD: 5.3368197

  Odds of: 1 vs 0

------------------------------------------------------------------------------
     ybinary |      b         z     P>|z|    bStdX    bStdY   bStdXY      SDofX
-------------+----------------------------------------------------------------
          x1 |   1.78923     9.815   0.000   3.5785   0.3353   0.6705     2.0000
          x2 |   1.17314     9.714   0.000   3.5194   0.2198   0.6595     3.0000
------------------------------------------------------------------------------
```

In the bivariate logistic regressions, the unstandardized coefficients for x1 and x2 are about .74 and .49 respectively; but when x1 and x2 are both in the equation, the coefficients are dramatically different, 1.79 and 1.17!

If we saw those kinds of changes in an OLS regression, we'd probably start thinking that suppressor effects were present, e.g. something like this might occur if x1 and x2 were negatively correlated while both had positive effects on y. But, since, in this example, x1 and x2 are uncorrelated, that is obviously not what is going on here. Rather, note how the standard deviation of Y* fluctuates from one logistic regression to the next; it is about 2.34 in each of the bivariate logistic regressions and 5.34 in the multivariate logistic regression.

Important! It is because the variance of Y* changes that the coefficients change so much when you go from one model to the next. In effect, the scaling of Y* is different in each model. By way of analogy, if in one OLS regression income was measured in dollars, and in another it was measured in thousands of dollars, the coefficients would be very different. See Appendix 1 for a discussion of what OLS would be like if it fixed the residual variance the same way that logistic regression does.

Compare this to the changes in the Y-Standardized and Fully-Standardized coefficients. The Y-Standardized coefficients for x1 and x2 are .3158 and .2095 in the bivariate regressions; in the multivariate regressions they are .3353 and .2198. Changes in the standardized coefficients are far less than the changes in the non-standardized coefficients. *If your goal is to compare coefficients across nested models, it is probably better to use Y-Standardized or Fully-Standardized coefficients.*

See Appendix 2 for an example using real data, and to also see how you can easily report y-standardized coefficients.

Question: In logistic regression, why does the variance of y* increase as you add more variables?

There are at least two ways of answering this. First, it increases because it has to. In OLS regression, SST (Total Sums of Squares) stays the same as you add more variables; but the Regression Sums of Squares is free to increase while the Error Sums of Squares makes a corresponding decrease. But, recall that, in logistic regression, $V(\varepsilon_{y*})$ is fixed at 3.29, i.e.

$$V(y^*) = V(\alpha + \sum X\beta) + V(\varepsilon_{y^*}) = V\left(\alpha + \sum X\beta\right) + \pi^2/3 = V\left(\alpha + \sum X\beta\right) + 3.29$$

Since the error variance can't go down, the explained variance (and hence the total variance) has to go up as you add more variables.

Optional. Second, the variance of y* changes because your estimation of the probability of success gets better and better as you add more variables, and better estimation in turn leads to more variability in the estimates. Suppose you had a single dichotomous IV, e.g. gender. All men would have the same probability of success (and hence the same logit) and all women would also share the same probability/logit. For example, men might have a 40% probability of success while women had a 50% chance.

As you add more variables, however, there can start to be more variability in the logits. So, among the men, after adding another variable the probabilities might now range between 35% and 45%, while for women the range might be 45% to 55%. Hence, instead of the probabilities only ranging from 40% to 50%, adding another variable could cause the probabilities to range from 35% to 55%, and the logits would vary more as well. If you are doing really really well, adding more variables might get you to the point where the probabilities ranged from .00001 to .9999, and the logits ranged from something like -10 to 10.

To put it another way: When relevant variables are missing, you will overestimate the probability of success for some (e.g. some men will have less than a 40% chance of success) while underestimating it for others (e.g. some women will have better than a 60% chance). As your model improves you will get more accurate estimates of the probability of success which in turn will result in greater variability in y*.

I am not sure, but I believe this second explanation helps to explain why y-standardization does not work perfectly, i.e. the variance of y* really is different across models (it isn't just that y* is getting scaled differently) so rescaling y* to have a variance of 1 does not fully solve the problem.

Another way of thinking about this: I've been highly critical of the use of standardized coefficients in OLS regression. But, in logistic regression, you are basically doing a different type of standardization: the variance of the residuals is fixed at $\pi^2/3$, or about 3.29. (Since you don't actually observe Y*, you have to identify its variance in some way.) Hence, logistic regression and other GLMs already have some of the problems of standardized coefficients inherently built into them. Given that an arbitrary type of standardization is already going on anyway, Y-Standardization or Full-Standardization may be superior for your purposes.

Question: How serious is the problem in practice? How can you avoid the problem?

- It won't be an issue at all if researchers don't attempt to compare coefficients across nested models. Indeed, it is very common to report model fit statistics (e.g. model chi-square, BIC, chi-square contrasts) for intermediate models, and to give coefficient estimates only for the final model. If you aren't going to focus much on changes in coefficients anyway, this is a very good strategy, and would probably be my first choice in most cases. Presenting a lot of unnecessary and potentially misleading coefficient estimates may do you more harm than good.
- If you do want to compare coefficients, it might not be that big of an issue if
    - V(y*) doesn't change that much from one model to the next.
    - Coefficients decline as you add variables rather than increase. In this case, you are actually underestimating the amount of decline. (Of course, this could be a problem too in that you may tend to overstate how important early variables are after adding controls.) Conversely, if coefficients increase as you add more variables, you have to be careful that any argument you want to make about suppressor effects is valid.
- Nonetheless, if you are going to compare coefficients, it seems best to use y-standardization, or at least check to make sure y-standardization would not change your conclusions.

My guess is that the problem doesn't come up that much in practice, partly because researchers using these methods seem less inclined to present detailed results for each model. Also, any comparisons that are made tend to focus on changes in statistical significance rather than changes in magnitudes of effects. But, as noted next, others disagree with my assessment.

Question: How well known is this problem? Are there any citations on this?

As noted before, Winship and Mare alluded to the need for y-standardization back in 1984, and they in turn were citing earlier work. Nonetheless, I don't think the problem is widely known and understood. If you want to compare coefficients across nested models and justify your use of y-standardization (and you want to quote something besides just a handout you found on the Internet) Mare briefly discusses the issue:

> When the error variance is fixed, it is also inappropriate to make within sample comparisons among the coefficients for a given covariate across equations with varying subsets of covariates. In this case, the total variance of the latent dependent variable and thus the scale of the estimated coefficients vary from model to model as a function of the different regressors that are included. Fixing the variance of the latent dependent variable avoids this problem. It does not, however, avoid the problems of comparison across samples and across dependent variables.

Source: *Response: Statistical Models of Educational Stratification—Hauser and Andrew's Models for School Transitions.* Robert D. Mare. Sociological Methodology 2006.

Carina Mood has written a very interesting article that appeared in the 2009 European Sociological Review entitled "Logistic Regression: Why we cannot do what we think we can do, and what we can do about it." She provides a more formal mathematical approach to the issues raised here and covers additional points. With regards to how prevalent the problem is, she states

> A look in any sociological journal that publishes quantitative research confirms that the problem of unobserved heterogeneity has escaped the attention of the large majority of users of logistic regression. LnOR and OR are interpreted as substantive effects, and it is common practice to compare coefficients across models within samples, and across samples, groups etc. just as in linear regression.

BUT, she doesn't provide a single citation to prove her point. However, a sociologist of education I talked to thinks she is right. S/he says

> This methodological issue is a really big one in the sociology of education and one that has mostly been ignored as far as I can tell. I bet you could find many examples of papers with these types of errors/issues…The problem in SOE is that we have so many outcomes that are truly categorical that we want to look at: dropping out, being in a particular class, getting certain grades, applying to college, etc. Things that are even more truly binary or categorical than "poverty status." And in SOE, folks are just not going to want to let go of the "comparing across models framework." For example, if we see that black students have twice the probability of being in a high track math class, we are then dying to find out how much of that is due to different factors, because that says something about the 'levers' we might pull to address this social problem… It would be great if there were some "rules of thumb" by which you could eyeball the extent of increase in coefficients due to scaling, which could be applied retroactively to already published analyses.

Of course, even if errors are widespread, it is not clear how serious they are. Is it really that horrible if, say, the coefficient for race is .618, when a corrected procedure would have put it at .600? I still don't think we have a good feel for how serious these problems are in practice.

I also like the way Mood phrases the problem. She basically says that *every model suffers from unobserved heterogeneity, i.e. every model has omitted variables.* For reasons I have already gone over, this affects the coefficients even when the omitted variables are uncorrelated with the variables in the model. *As a result, logistic regression coefficients are biased toward zero. As*

you add additional variables that affect the dependent variable the bias diminishes; but that means when you are comparing nested models, the model with fewer variables will be more biased toward zero than the model with more variables. Because the amount of bias differs across nested models, comparisons of coefficients can be misleading. She discusses the pros and cons of various ways for dealing with this, including y standardization.

Karlson, Holm and Breen have a working paper entitled "Comparing Regression Coefficients Between Models using Logit and Probit: A New Method." As of this writing (Sept. 11, 2011) it can be found at http://www.cser.dk/fileadmin/www.cser.dk/wp_003kbkkkjrb.pdf . I haven't carefully gone over their arguments, but if they are correct things are even more depressing than we previously thought, because, they claim, even Y standardization has problems. They do, however, propose some new solutions. Here is their abstract:

> Logit and probit models are widely used in empirical sociological research. However, the widespread practice of comparing the coefficients of a given variable across differently specified models does not warrant the same interpretation in logits and probits as in linear regression. Unlike in linear models, the change in the coefficient of the variable of interest cannot be straightforwardly attributed to the inclusion of confounding variables. The reason for this is that the variance of the underlying latent variable is not identified and will differ between models. We refer to this as the problem of rescaling. We propose a solution that allows researchers to assess the influence of confounding relative to the influence of rescaling, and we develop a test statistic that allows researchers to assess the statistical significance of both confounding and rescaling. We also show why y-standardized coefficients and average partial effects are not suitable for comparing coefficients across models. We present examples of the application of our method using simulated data and data from the National Educational Longitudinal Survey.

Conclusion: Often researchers present a hierarchy of models, e.g. they will estimate a model with x1-x3 included, then in a second model they will add x4-x6, then the third will add x7-x9, etc. As part of the discussion of the results, it might be noted how the effects of early variables decline or increase as additional variables are added, e.g. "The effect of race declines once income is controlled for." Such comparisons of coefficients are potentially misleading in a logistic regression; coefficient estimates can change, not just because the effect of a variable increases or decreases as others are controlled, but because V(Y*) is changing as new variables are added. Ergo, some of the things we are used to doing with metric coefficients in OLS regression are not legit when doing logistic regression. It may be best so simply avoid such comparisons and only present our preferred model. But, if we feel we must discuss how the effects of variables change as controls are added (and we may or may not be), Y-Standardized coefficients are probably betterm but even they may be problematic.

In OLS regression, I am not very fond of standardized coefficients. But, with logistic regression and other GLMs, some of the problems of standardized coefficients are inherently built in to the model, because these also do a type of standardization: the residuals are standardized to have a variance of 3.29. This has many advantages, but it also has drawbacks you need to be aware of. Depending on your purposes, you may find that a different type of standardization, e.g. Y-Standardization, is better. (Of course, then you have to explain to the reader what those are!) In any event, you should be aware of the potential pitfalls, e.g. your discussion should be careful about making comparisons of coefficients across models that may not be valid.

Appendix 1

RWLS as an alternative to OLS (or why we should be grateful OLS isn't more like logistic regression). In logistic regression, the variance of ε_{y^*} is fixed at 3.29. In OLS, the Total Sums of Squares stays the same as you add variables, but the regression and error sums of squares can vary. Suppose (perhaps out of some insane desire for consistency) we had a variation of OLS which did something similar to logistic regression, e.g. it standardized Y so that the mean square error (MSE) in a regression was always 3.29. We'll call this alternate-universe version of ordinary least squares RWLS (Rich Williams Least Squares). How would regression coefficients behave under such an alternate method? The following examples show what would happen.

```
. use https://www3.nd.edu/~rwilliam/statafiles/standardized.dta
. quietly reg y x1
. gen double ystar = y / sqrt((e(rss)/e(df_r))) * sqrt(3.29)
. reg ystar x1, beta

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  398.40
       Model |  1310.73621     1  1310.73621           Prob > F      =  0.0000
    Residual |     1638.42   498        3.29           R-squared     =  0.4444
-------------+------------------------------           Adj R-squared =  0.4433
       Total |  2949.15621   499  5.91013269           Root MSE      =  1.8138

------------------------------------------------------------------------------
       ystar |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
          x1 |   .8103589   .0405992    19.96   0.000                .6666667
       _cons |   4.65e-07   .0811172     0.00   1.000                       .
------------------------------------------------------------------------------

. quietly reg y x2
. replace ystar = y / sqrt((e(rss)/e(df_r))) * sqrt(3.29)

(500 real changes made)

. reg ystar x2, beta

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  398.40
       Model |  1310.73549     1  1310.73549           Prob > F      =  0.0000
    Residual |     1638.42   498        3.29           R-squared     =  0.4444
-------------+------------------------------           Adj R-squared =  0.4433
       Total |  2949.15549   499  5.91013125           Root MSE      =  1.8138

------------------------------------------------------------------------------
       ystar |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
          x2 |   .5402391   .0270661    19.96   0.000                .6666666
       _cons |   4.28e-07   .0811172     0.00   1.000                       .
------------------------------------------------------------------------------

. quietly reg y x1 x2
. replace ystar = y / sqrt((e(rss)/e(df_r))) * sqrt(3.29)

(500 real changes made)
```

```
. reg ystar x1 x2, beta

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) = 1988.00
       Model |  13081.0321     2  6540.51606           Prob > F      =  0.0000
    Residual |     1635.13   497        3.29           R-squared     =  0.8889
-------------+------------------------------           Adj R-squared =  0.8884
       Total |  14716.1621   499  29.4913069           Root MSE      =  1.8138

------------------------------------------------------------------------------
       ystar |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
          x1 |   1.810197   .0405992    44.59   0.000                 .6666667
          x2 |   1.206798   .0270661    44.59   0.000                 .6666666
       _cons |   9.95e-07   .0811172     0.00   1.000                        .
------------------------------------------------------------------------------
```

In the RWLS bivariate regressions, the unstandardized coefficients for x1 and x2 are about .81 and .54 respectively; but when both x1 and x2 are both in the equation, the coefficients are dramatically different, 1.81 and 1.21 (unlike OLS, where they didn't change at all). Further, the variance of ystar (as shown by the MS Total) is about 5.91 in each of the bivariate regressions but zooms to 29.49 in the multivariate regression. However, the standardized coefficients are the same throughout. In short, *if OLS was more like logistic regression, where the error variance was fixed instead of free to vary, we'd see the same sort of oddities in the parameter estimates as we went from one model to the next as we did with logistic regression*.

I've often noted the evils of standardized coefficients in OLS. But, if I had to choose between RWLS and standardized coefficients, I might grudgingly go with standardized coefficients – at least they don't create the same kind of moving target with V(Y) that RWLS does. Luckily, we don't have to use RWLS! (But unfortunately, I'll have to keep looking for a statistical technique I can name after myself – this one is doomed to either forever live in infamy or else be quickly forgotten.)

Appendix 2: Reporting Y-Standardized Coefficients

The `estadd` routine available from SSC makes it possible to easily report y-standardized coefficients. `spost9` also needs to be installed.

```
. webuse nhanes2f, clear
. quietly logit diabetes black
. quietly estadd listcoef, std
. est store m1
. quietly logit diabetes black age
. quietly estadd listcoef, std
. est store m2
```

```
. esttab m1 m2, pr2

--------------------------------------------
                      (1)             (2)
                 diabetes        diabetes
--------------------------------------------
diabetes
black               0.610***        0.718***
                   (4.95)          (5.66)

age                                 0.0595***
                                   (15.94)

_cons              -3.063***       -6.324***
                  (-60.84)        (-27.32)
--------------------------------------------
N                   10335           10335
pseudo R-sq         0.005           0.093
--------------------------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

. esttab m1 m2, main(b_ys) pr2

--------------------------------------------
                      (1)             (2)
                 diabetes        diabetes
--------------------------------------------
diabetes
black               0.334***        0.343***
                   (4.95)          (5.66)

age                                 0.0285***
                                   (15.94)

_cons                     ***             ***
                  (-60.84)        (-27.32)
--------------------------------------------
N                   10335           10335
pseudo R-sq         0.005           0.093
--------------------------------------------
b_ys coefficients; t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

The first table gives you the regular coefficients; the second table gives you the y-standardized coefficients. Note that the first table makes it look like the effect of black increases almost 18 percent once age is added to the model; whereas in the 2nd table, using y-standardized coefficients, the effect of black changes hardly at all. This is primarily because the latent SD of Y* jumped from 1.82 in the first model to 2.09 in the second, or about 15 percent. There is also a very slight suppressor effect present because black and age are negatively correlated (-.0321) while both have a positive effect on diabetes.