Comparing Logit and Probit Coefficients Across Groups: Problems, Solutions, and Problems with the Solutions

Richard Williams
Notre Dame Sociology
rwilliam@ND.Edu
https://www3.nd.edu/~rwilliam
Last revised November 5, 2024

- We often want to compare the effects of variables across groups, e.g. we want to see if the effect of education is the same for men as it is for women
- Both OLS and logistic regression assume that error variances are the same for both groups

- When that assumption is violated in OLS, the consequences are often minor: standard errors and significance tests are a bit off but coefficients remain unbiased.

- But when a binary or ordinal regression model incorrectly assumes that error variances are the same for all cases, the standard errors are wrong and (unlike OLS regression) the parameter estimates are wrong too.

- We often think that the observed binary or ordinal variable y is a collapsed version of a latent continuous unobserved variable y*.
- Because y* is unobserved, its metric has to be fixed in some way. This is typically done by scaling y* so that its residual variance is $\pi^2/3 = 3.29$.
- But this creates problems similar to those encountered when analyzing standardized coefficients in OLS
  - unless the residual variance really is the same in both groups (i.e. errors are homoskedastic) the coefficients will be scaled differently and will *not* be comparable.

- As Hoetker (2004, p. 17) notes, “in the presence of even fairly small differences in residual variation, naive comparisons of coefficients [across groups] can indicate differences where none exist, hide differences that do exist, and even show differences in the opposite direction of what actually exists.”
- **Explanation.** Suppose that y* were observed, but our estimation procedure continued to standardize the variable by fixing its residual variance at 3.29. How would differences in residual variability across groups affect the estimated coefficients?
  - In the examples, the coefficients for the residuals reflect the differences in residual variability across groups.
  - Any residual that does not have a coefficient attached to it is assumed to already have a variance of 3.29

CASE 1: UNDERLYING ALPHAS ARE EQUAL, RESIDUAL VARIANCES DIFFER

	GROUP 0	GROUP 1
TRUE COEFFICIENTS	$y_i^* = x_{i1} + x_{i2} + x_{i3} + \varepsilon_i$	$y_i^* = x_{i1} + x_{i2} + x_{i3} + 2\varepsilon_i$
STANDARDIZED COEFFICIENTS	$y_i^* = x_{i1} + x_{i2} + x_{i3} + \varepsilon_i$	$\frac{y_i^*}{2} = \frac{x_{i1}}{2} + \frac{x_{i2}}{2} + \frac{x_{i3}}{2} + \varepsilon_i$

In Case 1, the true coefficients all equal 1 in both groups. But, because the residual variance is twice as large for group 1 as it is for group 0, the standardized βs are only half as large for group 1 as for group 0. *Naive comparisons of coefficients can indicate differences where none exist*.

Case 2: Underlying alphas differ, residual variances differ

	Group 0	Group 1
True Coefficients	$y_i^* = x_{i1} + x_{i2} + x_{i3} + \varepsilon_i$	$y_i^* = 2x_{i1} + 2x_{i2} + 2x_{i3} + 2\varepsilon_i$
Standardized Coefficients	$y_i^* = x_{i1} + x_{i2} + x_{i3} + \varepsilon_i$	$\frac{y_i^*}{2} = x_{i1} + x_{i2} + x_{i3} + \varepsilon_i$

In Case 2, the true coefficients are twice as large in group 1 as in group 0. But, because the residual variances also differ, the standardized βs for the two groups are the same. Differences in residual variances obscure the differences in the underlying effects. *Naive comparisons of coefficients can hide differences that do exist.*

Case 3: Underlying alphas differ, residual variances differ even more

	Group 0	Group 1
True Coefficients	$y_i^* = x_{i1} + x_{i2} + x_{i3} + \varepsilon_i$	$y_i^* = 2x_{i1} + 2x_{i2} + 2x_{i3} + 3\varepsilon_i$
Standardized Coefficients	$y_i^* = x_{i1} + x_{i2} + x_{i3} + \varepsilon_i$	$\frac{y_i^*}{3} = \frac{2}{3}x_{i1} + \frac{2}{3}x_{i2} + \frac{2}{3}x_{i3} + \varepsilon_i$

In Case 3, the true coefficients are again twice as large in group 1 as in group 0. But, because of the large differences in residual variances, the standardized βs are smaller for group 0 than group 1. Differences in residual variances make it look like the Xs have smaller effects on group 1 when really the effects are larger. *Naive comparisons of coefficients can even show differences in the opposite direction of what actually exists*.

Example: Allison's (1999) model for group comparisons

- Allison (Sociological Methods and Research, 1999) analyzes a data set of 301 male and 177 female biochemists.

- Allison uses logistic regressions to predict the probability of promotion to associate professor.

Table 1: Results of Logit Regressions Predicting Promotion to Associate Professor for Male and Female Biochemists (Adapted from Allison 1999, p. 188)

	Men		*Women*		*Ratio of*	*Chi-Square*
Variable	*Coefficient*	*SE*	*Coefficient*	*SE*	*Coefficients*	*for Difference*
Intercept	-7.6802***	.6814	-5.8420***	.8659	.76	2.78
Duration	1.9089***	.2141	1.4078***	.2573	.74	2.24
Duration squared	-0.1432***	.0186	-0.0956***	.0219	.67	2.74
Undergraduate selectivity	0.2158***	.0614	0.0551	.0717	.25	2.90
Number of articles	*0.0737****	*.0116*	*0.0340***	*.0126*	*.46*	*5.37**
Job prestige	-0.4312***	.1088	-0.3708*	.1560	.86	0.10
Log likelihood	-526.54		-306.19			
Error variance	3.29		3.29			

**p < .05, **p < .01, *** p < .001*

- As his Table 1 shows, the effect of number of articles on promotion is about twice as great for males (.0737) as it is for females (.0340).
- If accurate, this difference suggests that men get a greater payoff from their published work than do females, “a conclusion that many would find troubling” (Allison 1999:186).
- BUT, Allison warns, women may have more heterogeneous career patterns, and unmeasured variables affecting chances for promotion may be more important for women than for men.
  - Put another way, the error variance for women may be greater than the error variance for men
  - This corresponds to the Case I we presented earlier.
  - Unless the residual variability is identical across populations, the standardization of coefficients for each group will also differ.

Allison's solution for the problem

- Ergo, in his Table 2, Allison adds a parameter to the model he calls delta. Delta adjusts for differences in residual variation across groups.

Table 2: Logit Regressions Predicting Promotion to Associate Professor for Male and Female Biochemists, Disturbance Variances Unconstrained (Adapted from Allison 1999, p. 195)

	All Coefficients Equal		*Articles Coefficient Unconstrained*	
Variable	*Coefficient*	*SE*	*Coefficient*	*SE*
Intercept	-7.4913***	.6845	-7.3655***	.6818
Female	-0.93918**	.3624	-0.37819	.4833
Duration	1.9097***	.2147	1.8384***	.2143
Duration squared	-0.13970***	.0173	-0.13429***	.01749
Undergraduate selectivity	0.18195**	.0615	0.16997***	.04959
Number of articles	0.06354***	.0117	0.07199***	.01079
Job prestige	-0.4460***	.1098	-0.42046***	.09007
δ	*-0.26084**	*.1116*	*-0.16262*	*.1505*
Articles x Female			*-0.03064*	*.0173*
Log likelihood	-836.28		-835.13	

**p < .05, **p < .01, *** p < .001*

- The delta-hat coefficient value –.26 in Allison's Table 2 (first model) tells us that the standard deviation of the disturbance variance for men is 26 percent lower than the standard deviation for women.
  - This implies women have more variable career patterns than do men, which causes their coefficients to be lowered relative to men when differences in variability are not taken into account, as in the original logistic regressions.

- Allison's final model shows that the interaction term for Articles x Female is NOT statistically significant

- Allison concludes "The apparent difference in the coefficients for article counts in Table 1 does not necessarily reflect a real difference in causal effects. It can be readily explained by differences in the degree of residual variation between men and women."

Problems with Allison's Approach

- Williams (2009) noted various problems with Allison's approach

- Allison's test has difficulty distinguishing between cross-group differences in residual variability & differences in coefficients. (I won't explain why here; you can read the paper.)

- Also, Allison's approach only allows for a single categorical variable in the variance equation. The sources of heteroskedasticity can be more complex than that; more variables may be involved, & some of these may be continuous
- Keele & Park (2006) show that a mis-specificied variance equation, e.g. one in which relevant variables are omitted, can actually be worse than having no variance equation at all.

- Finally, Allison's method only works with a dichotomous dependent variable
  - Models with binary dvs that allow for heteroskedasticity can be difficult to estimate
  - Ordinal dependent variables contain more information about Y*
- Williams (2009, 2010) therefore proposed a more powerful alternative

A Broader Solution: Heterogeneous Choice Models

- Heterogeneous choice/ location-scale models explicitly specify the determinants of heteroskedasticity in an attempt to correct for it.

- These models are also useful when the variability of underlying attitudes is itself of substantive interest.

The Heterogeneous Choice (aka Location-Scale) Model

- Can be used for binary or ordinal models
- Two equations, choice & variance
- Binary case :

$$\Pr(y_i = 1) = g\left(\frac{x_i\beta}{\exp(z_i\gamma)}\right) = g\left(\frac{x_i\beta}{\exp(\ln(\sigma_i))}\right) = g\left(\frac{x_i\beta}{\sigma_i}\right)$$

- Allison's model with delta is actually a special case of a heterogeneous choice model, where the dependent variable is a dichotomy and the variance equation includes a single dichotomous variable that also appears in the choice equation.
- Allison's results can easily be replicated with the user-written routine oglm (Williams, 2009, 2010)

```
. * oglm replication of Allison's Table 2, Model 2 with interaction added:
. use "http://www.indiana.edu/~jslsoc/stata/spex_data/tenure01.dta", clear
(Gender differences in receipt of tenure (Scott Long 06Jul2006))
. keep if pdasample
(148 observations deleted)
. oglm tenure female year yearsq select articles prestige f_articles, het(female)

Heteroskedastic Ordered Logistic Regression       Number of obs   =       2797
                                                  LR chi2(8)      =     415.39
                                                  Prob > chi2     =     0.0000
Log likelihood = -835.13347                       Pseudo R2       =     0.1992

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
tenure       |
      female |  -.3780597   .4500207    -0.84   0.401    -1.260084    .5039646
        year |   1.838257   .2029491     9.06   0.000     1.440484     2.23603
      yearsq |  -.1342828    .017024    -7.89   0.000    -.1676492   -.1009165
      select |   .1699659   .0516643     3.29   0.001     .0687057    .2712261
    articles |   .0719821   .0114106     6.31   0.000     .0496178    .0943464
    prestige |  -.4204742   .0961206    -4.37   0.000    -.6088671   -.2320813
  f_articles |  -.0304836   .0187427    -1.63   0.104    -.0672185    .0062514
-------------+----------------------------------------------------------------
lnsigma      |
      female |   .1774193   .1627087     1.09   0.276     -.141484    .4963226
-------------+----------------------------------------------------------------
       /cut1 |   7.365285   .6547121    11.25   0.000     6.082073    8.648497
------------------------------------------------------------------------------
. display "Allison's delta = " (1 - exp(.1774193)) / exp(.1774193)
-.16257142
```

- As Williams (2009) notes, there are important advantages to turning to the broader class of heterogeneous choice models that can be estimated by oglm
  - Dependent variables can be ordinal rather than binary. This is important, because ordinal vars have more information and hence lead to better estimation
  - The variance equation need not be limited to a single binary grouping variable, which (hopefully) reduces the likelihood that the variance equation will be mis-specified

- Williams (2010) also notes that, even if the researcher does not want to present a heterogenous choice model, estimating one can be useful from a diagnostic standpoint
  - Often, the appearance of heteroskedasticity is actually caused by other problems in model specification, e.g. variables are omitted, variables should be transformed (e.g. logged), squared terms should be added
  - Williams (2010) shows that the heteroskedasticity issues in Allison's models go away if articles^2 is added to the model

```
. use "http://www.indiana.edu/~jslsoc/stata/spex_data/tenure01.dta", clear
(Gender differences in receipt of tenure (Scott Long 06Jul2006))
. keep if pdasample
(148 observations deleted)
. * hetero effect becomes insignificant when articles^2 is added to model
. oglm tenure i.female year c.year#c.year select articles prestige c.articles#c.articles, het(i.female)

Heteroskedastic Ordered Logistic Regression       Number of obs   =       2797
                                                  LR chi2(8)      =     440.07
                                                  Prob > chi2     =     0.0000
Log likelihood = -822.79102                       Pseudo R2       =     0.2110

------------------------------------------------------------------------------
             tenure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------+---------------------------------------------------------
tenure              |
           1.female |  -.5612041   .2926285    -1.92   0.055    -1.134746    .0123373
               year |   1.739173   .1933258     9.00   0.000     1.360261    2.118084
                    |
      c.year#c.year |  -.1265911   .0162677    -7.78   0.000    -.1584751    -.094707
                    |
             select |   .1710519   .0504271     3.39   0.001     .0722167    .2698872
           articles |   .1533931   .0244003     6.29   0.000     .1055694    .2012168
           prestige |   -.454951   .0936162    -4.86   0.000    -.6384354   -.2714666
                    |
c.articles#c.articles |  -.0026412   .0007213    -3.66   0.000     -.004055   -.0012274
--------------------+---------------------------------------------------------
lnsigma             |
           1.female |    .141633   .1377843     1.03   0.304    -.1284193    .4116853
--------------------+---------------------------------------------------------
             /cut1 |    7.40805    .648316    11.43   0.000     6.137374    8.678726
------------------------------------------------------------------------------
```

```
. * You don't near the female*articles interaction terms either

. oglm tenure i.female year c.year#c.year select articles prestige c.articles#c.articles
i.female#(c.articles c.articles#c.articles)

Ordered Logistic Regression                     Number of obs   =       2797
                                                LR chi2(9)      =     439.05
                                                Prob > chi2     =     0.0000
Log likelihood =  -823.3041                     Pseudo R2       =     0.2105

------------------------------------------------------------------------------
             tenure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------+---------------------------------------------------------
           1.female |  -.2588495   .3094906    -0.84   0.403    -.8654399    .3477409
               year |   1.649057   .1651954     9.98   0.000      1.32528    1.972834
                    |
      c.year#c.year |    -.11984   .0142338    -8.42   0.000    -.1477378   -.0919423
                    |
             select |   .1583254   .0466906     3.39   0.001     .0668136    .2498372
           articles |   .1492724   .0295934     5.04   0.000     .0912703    .2072745
           prestige |  -.4386555   .0898018    -4.88   0.000    -.6146637   -.2626472
                    |
 c.articles#c.articles |  -.0025455   .0009236    -2.76   0.006    -.0043558   -.0007352
                    |
    female#c.articles |
                  1 |   -.007599   .0450029    -0.17   0.866    -.0958031    .0806051
                    |
female#c.articles#c.articles |
                  1 |   .0001025   .0013542     0.08   0.940    -.0025517    .0027568
--------------------+---------------------------------------------------------
               /cut1 |   7.091958   .5479358    12.94   0.000     6.018024    8.165893
------------------------------------------------------------------------------
```

Problems with heterogeneous choice models

- Models can be difficult to estimate, although this is generally less problematic with ordinal variables
- While you have more flexibility when specifying the variance equation, a mis-specified equation can still be worse than no equation at all
- But the most critical problem of all may be…

Problem: Radically different interpretations are possible

- An issue to be aware of with heterogeneous choice models is that radically different interpretations of the results are possible
  - Hauser and Andrew (2006), for example, proposed a seemingly different model for assessing differences in the effects of variables across groups (where in their case, the groups were different educational transitions)
  - They called it the *logistic response model with proportionality constraints* (LRPC):

$$\log_e\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{j0} + \lambda_j \sum_k \beta_k X_{ijk}$$

- Instead of having to estimate a different set of coefficients for each group/transition, you estimate a single set of coefficients, along with one λj proportionality factor for each group/ transition (λ_1 is constrained to equal 1)
  - The proportionality constraints would hold if, say, the coefficients for the 2nd group were all 2/3 as large as the corresponding coefficients for the first group, the coefficients for the 3rd group were all half as large as for the first group, etc.

Models compared

$$\Pr(y_i = 1) = g\left(\frac{x_i\beta}{\exp(z_i\gamma)}\right) = g\left(\frac{x_i\beta}{\exp(\ln(\sigma_i))}\right) = g\left(\frac{x_i\beta}{\sigma_i}\right)$$

$$\log_e\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{j0} + \lambda_j \sum_k \beta_k X_{ijk}$$

- Hauser & Andrew note, however, that “one cannot distinguish empirically between the hypothesis of uniform proportionality of effects across transitions and the hypothesis that group differences between parameters of binary regressions are artifacts of heterogeneity between groups in residual variation.” (p. 8)
- Williams (2010) showed that, even though the rationales behind the models are totally different, the heterogeneous choice models estimated by oglm produce identical fits to the LRPC models estimated by Hauser and Andrew; simple algebra converts one model’s parameters into the other’s
- Williams further showed that Hauser & Andrew’s software produced the exact same coefficients that Allison’s software did when used with Allison’s data

```
. * Hauser & Andrew's original LRPC program
. * Code has been made more efficient and readable,
. * but results are the same.  Note that it
. * actually estimates and reports
. * lambda - 1 rather than lamba.
. program define lrpc02
  1.         tempvar theta
  2.         version 8
  3.         args lnf  intercepts lambdaminus1 betas
  4.         gen double `theta' = `intercepts' + `betas' + (`lambdaminus1' * `betas')
  5.         quietly replace `lnf' = ln(exp(`theta')/(1+exp(`theta'))) if $ML_y1==1
  6.         quietly replace `lnf' = ln(1/(1+exp(`theta'))) if $ML_y1==0
  7. end
```

```
. * Hauser & Andrews original LRPC parameterization used with Allison's data
. * Results are identical to Allison's Table 2, Model 1
. ml model lf lrpc02 ///
>         (intercepts: tenure = male female, nocons) ///
>         (lambdaminus1: female, nocons) ///
>         (betas: year yearsq select articles prestige, nocons), max nolog
. ml display

                                                  Number of obs   =       2797
                                                  Wald chi2(2)    =     180.60
Log likelihood = -836.28235                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
      tenure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
intercepts   |
        male |  -7.490506   .6596634   -11.36   0.000    -8.783422   -6.197589
      female |  -6.230958   .6205863   -10.04   0.000    -7.447285   -5.014631
-------------+----------------------------------------------------------------
lambdaminus1 |
      female |  -.2608325   .1080502    -2.41   0.016    -.4726069   -.0490581
-------------+----------------------------------------------------------------
betas        |
        year |   1.909544   .1996937     9.56   0.000     1.518151    2.300936
      yearsq |  -.1396868   .0169425    -8.24   0.000    -.1728935   -.1064801
      select |   .1819201   .0526572     3.45   0.001     .0787139    .2851264
    articles |   .0635345    .010219     6.22   0.000     .0435055    .0835635
    prestige |  -.4462074    .096904    -4.60   0.000    -.6361357   -.2562791
------------------------------------------------------------------------------
```

- But, the theoretical concerns that motivated their models and programs lead to radically different interpretations of the results.
  - According to Allison's theory (and the theory behind the heterogeneous choice model) apparent differences in effects between men and women are an artifact of differences in residual variability.
  - Once these differences are taken into account, there is no significant difference in the effect of articles across groups, implying there is no gender inequality in the tenure process.

- Someone looking at these exact same numbers from the viewpoint of the LRPC, however, would conclude that the effect of articles (and every other variable for that matter) is 26 percent smaller for women than it is men.
- Those who believed that the LRPC was the theoretically correct model would likely conclude that there is substantial gender inequality in the tenure promotion process.
- For any given problem, strong substantive arguments might be made for one perspective or the other.
- Researchers using any of these models should realize, however, that there is often if not always a radically different interpretation that, empirically, fits the data just as well.

Long's solution

- Long (2009) looks at these same sorts of problems, but proposes a different analytical approach. He says
  - "An alternative approach [to Allison]… uses predicted probabilities. Since predicted probabilities are unaffected by residual variation, tests of the equality of predicted probabilities across groups can be used for group comparisons without assuming the equality of the regression coefficients of some variables… Testing the equality of predicted probabilities requires multiple tests since group differences in predictions vary with the levels of the variables in the model."

- Long's approach lets all coefficients differ by group. In the following example he uses interaction terms so that the male and female coefficients can freely differ.
- He then estimates marginal effects for the gender variable across a range of values for # of articles, to see whether and how the predicted values for men and women differ

A simple example of Long's technique

```
. use "http://www.indiana.edu/~jslsoc/stata/spex_data/tenure01.dta", clear
(Gender differences in receipt of tenure (Scott Long 06Jul2006))
. keep if year <= 10
(148 observations deleted)
. * Basic model - articles only
. logit tenure articles i.male i.male#c.articles, nolog

Logistic regression                               Number of obs   =       2797
                                                  LR chi2(3)      =     121.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -982.04029                       Pseudo R2       =     0.0583

------------------------------------------------------------------------------
      tenure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    articles |   .0471351   .0104974     4.49   0.000     .0265605    .0677097
      1.male |  -.2198428   .1853876    -1.19   0.236    -.5831959    .1435102
             |
male#c.articles |
          1  |   .0552514   .0148436     3.72   0.000     .0261585    .0843444
             |
       _cons |  -2.501162    .140056   -17.86   0.000    -2.775667   -2.226657
------------------------------------------------------------------------------
```

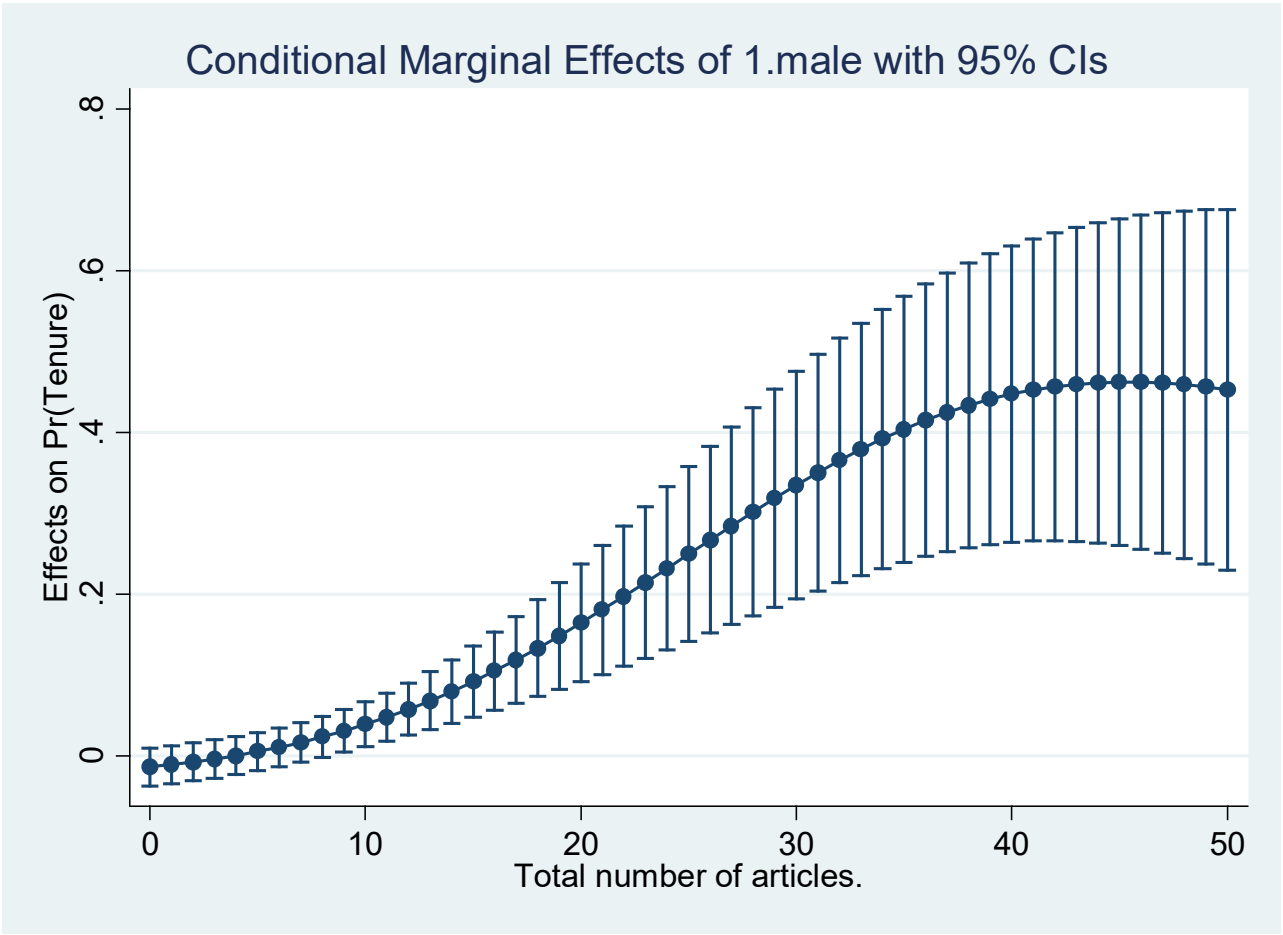
```
. margins, dydx(male) at(articles=(0(1)50)) vsquish

Conditional marginal effects                      Number of obs   =       2797
Model VCE    : OIM

Expression   : Pr(tenure), predict()
dy/dx w.r.t. : 1.male
1._at        : articles        =           0
2._at        : articles        =           1
 [output deleted]
51._at       : articles        =          50

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.male       |
         _at |
          1  |  -.0140315   .0120717    -1.16   0.245    -.0376916    .0096285
          2  |  -.0111948   .0120559    -0.93   0.353    -.0348239    .0124343
 [output deleted]
         50  |   .4562794   .1118036     4.08   0.000     .2371485    .6754104
         51  |   .4527383   .1139979     3.97   0.000     .2293066      .67617
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.

. marginsplot
```


Conditional Marginal Effects of 1.male with 95% CIs
Effects on Pr(Tenure)
0
.2
.4
.6
.8
0
10
20
30
40
50
Total number of articles.

- This simple example shows that the predicted probabilities of tenure for men and women differ little for small numbers of articles; indeed the differences are not even statistically significant for 8 articles or less.
- The differences become greater as the number of articles increases. For example, a women with 40 articles is predicted to be 45 percent less likely to get tenure than a man with 40 articles.

- The analyses can be further extended by adding more variables to the model, and/or by doing various subgroup analyses, e.g. comparing women at high-prestige universities with men at high prestige Universities
- As Long says, this can lead to “more complex conclusions on how groups differ in the effect of a variable.”
- If you are lucky, the differences in predicted probabilities may disappear altogether, e.g. variables added to the model may be able to account for the initially observed group differences.
- But if they don’t…

Critique of Long

- The predictive margins produced by Long's approach might be seen as a sort of high-tech descriptives. They illustrate the predicted differences between groups after controlling for other variables.
- Description can be very useful. In this case we see that the predicted probabilities of tenure differ dramatically by gender and the number of articles published.
- Once such differences in predicted probabilities are discovered, policy makers may decide that some sort of corrective action should be considered.

- At the same time, Long's approach may be frustrating because it doesn't try to explain why the differences exist.
  - Are the differences due to the fact that men are rewarded more for the articles they publish?
  - Or, are they due to the fact that residual variability differs by gender? Perhaps women's careers are disrupted more by family or other matters.
  - Long's approach lets all coefficients differ by group, rather than try to determine which variable effects are different in each group.

- From a policy standpoint, we would like to know what is causing these observed differences in predicted probabilities
  - If it is because women are rewarded less for each article they write, we may want to examine if women's work is not being evaluated fairly
  - If it is because of differences in residual variability, we may want to further examine why that is. For example, if family obligations create more career hurdles for women then they do men, how can we make the workplace more family-friendly?
  - But if we do not know what is causing the differences, we aren't even sure where to start if we want to eliminate them.
  - In short, Long's approach using marginal effects lets us see where differences exist across groups but does not try to explain what causes them.

- Long defends his approach by arguing:
  - For many things, like his report on women in science for the NAS, predictions were of much more interest than was the slope of articles or unobserved heterogeneity.
  - using other information, e.g. on the history of women in science, may resolve issues far more effectively than the types of assumptions that are needed to be able to disentangle differences in coefficients and unobserved heterogeneity
  - there are times when predictive margins provide more insights than simple answers to yes no hypotheses. For example, there can be cases where, overall the lines for men and women are the same (can't reject they are equal), yet they differ significantly when testing equality at a particular case. Both are valid, but overreliance on one, omnibus test is not a good thing in general.

- Further, as we have seen, when we try to explain group differences, the coefficients can be interpreted in radically different ways.
  - Two researchers could look at the exact same set of results, and one could conclude that coefficients differ across groups while another could say that it is residual variability that differs.
- Given such ambiguity, some might argue that you should settle for description and not strive for explanation (at least not with the current data).

- Others might argue that you should go with the model that you think makes most theoretical sense, while acknowledging that alternative interpretations of the results are possible.

- At this point, it is probably fair to say that the descriptions of the problem may be better, or at least more clear-cut, than the various proposed solutions.

- Long & Mustillo 2018 (https://doi.org/10.1177/0049124118799374) and Mize, Doan, & Long 2019 (https://journals.sagepub.com/doi/10.1177/0081175019852763) have further refined Long's 2009 arguments.
- MDL have further shown a way to test whether marginal effects significantly differ across groups.
- Still, while approaches may have gotten more sophisticated, the limitations of using marginal effects to explain why groups differ remain.

Selected References

- Allison, Paul. 1999. Comparing Logit and Probit Coefficients Across Groups. *Sociological Methods and Research* 28(2): 186-208.
- Hauser, Robert M. and Megan Andrew. 2006. Another Look at the Stratification of Educational Transitions: The Logistic Response Model with Partial Proportionality Constraints. *Sociological Methodology* 36(1):1-26.
- Hoetker, Glenn. 2004. Confounded Coefficients: Extending Recent Advances in the Accurate Comparison of Logit and Probit Coefficients Across Groups. Working Paper, October 22, 2004. Retrieved September 27, 2011 (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=609104)
- Keele, Luke and David K. Park. 2006. Difficult Choices: An Evaluation of Heterogeneous Choice Models. Working Paper, March 3, 2006. Retrieved March 21, 2006 (https://www3.nd.edu/~rwilliam/oglm/ljk-021706.pdf)
- Karlson, Kristian B., Anders Holm and Richard Breen. 2011. Comparing Regression Coefficients between Same-Sample Nested Models using Logit and Probit: A New Method. Forthcoming in *Sociological Methodology*.
- Kohler, Ulrich, Kristian B. Carlson and Anders Holm. 2011. Comparing Coefficients of nested nonlinear probability models. Forthcoming in *The Stata Journal*.
- Long, J. Scott. 2009. Group comparisons in logit and probit using predicted probabilities. Working Paper, June 25, 2009. Retrieved September 27, 2011 (http://www.indiana.edu/~jslsoc/files_research/groupdif/groupwithprobabilities/groups-with-prob-2009-06-25.pdf)
- Long, J. Scott and Jeremy Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata*, 2nd Edition. College Station, Texas: Stata Press.
- Williams, Richard. 2009. Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients across Groups. *Sociological Methods & Research* 37(4): 531-559. A pre-publication version is available at https://www3.nd.edu/~rwilliam/oglm/RW_Hetero_Choice.pdf.
- Williams, Richard. 2010. Fitting Heterogeneous Choice Models with oglm. *The Stata Journal* 10(4):540-567. Available at http://www.stata-journal.com/article.html?article=st0208.

For more information, see:

https://www3.nd.edu/~rwilliam