Group Comparisons and Other Issues in Interpreting Models for Categorical Outcomes Using Stata and SPost

Scott Long, Indiana University

July 2006

1. Interpretation of regression models for categorical outcomes.
2. Focus on using predictions rather than coefficients due to nonlinearity of model.
3. Begin with a simple example predicting tenure.
4. Extend ideas to methods for comparing groups.
5. Illustrate how Stata's programming features can be used in do files using Stata with SPost (with Jeremy Freese).

Logit

$$\begin{aligned}\Pr(y=1\mid\mathbf{x}) &= \Lambda\left(\beta_0+\beta_x x+\beta_z z\right)\\ &= \frac{\exp\left(\beta_0+\beta_x x+\beta_z z\right)}{1+\exp\left(\beta_0+\beta_x x+\beta_z z\right)}\end{aligned}$$

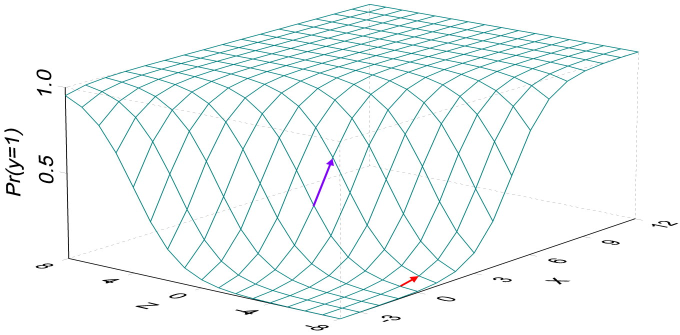
Probit

$$\Pr(y=1\mid\mathbf{x}) = \Phi\left(\beta_0+\beta_x x+\beta_z z\right)$$

Methods apply to other cross-sectional models (`regress`, `ologit`, `oprobit`, `mlogit`, `asmprobit`, `mprobit`, `nbreg`, `poisson`, `zip`, `zinb`).

Interpretation of regression models with Stata and SPost

The challenge of nonlinearity



Example: gender differences in tenure

Descriptive statistics

1. Binary outcome is receipt of tenure for 301 male biochemists and 177 female biochemists.
2. Each observation is a person-year in rank (hence, an event history model).

```
. use tenure01, clear

. vardesc tenure female year yearsq select articles prestige presthi

Var            Mean    StdDev   Minimum   Maximum   Description
tenure         0.12     0.329      0.00      1.00   Is tenured?
female         0.38     0.486      0.00      1.00   Scientist is female?
year           4.33     3.090      1.00     22.00   Years in rank.
yearsq        28.29    44.181      1.00    484.00   Years in rank squared.
select         4.97     1.434      1.00      7.00   Selectivity of bachelor's.
articles       7.21     6.745      0.00     73.00   Total number of articles.
prestige       2.63     0.771      0.65      4.80   Prestige of department.
presthi        0.05     0.208      0.00      1.00   Prestige is 4 or higher?

N = 2945
```

Example: gender differences in tenure

Estimates from logit

```
. logit tenure female year yearsq select articles presthi, nolog

Logistic regression                               Number of obs   =       2945
                                                  LR chi2(6)      =     336.43
                                                  Prob > chi2     =     0.0000
Log likelihood = -931.36045                       Pseudo R2       =     0.1530

------------------------------------------------------------------------------
      tenure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.3735409   .1270093    -2.94   0.003    -.6224745   -.1246073
        year |    .932456    .084743    11.00   0.000     .7663627    1.098549
      yearsq |  -.0538009   .0060204    -8.94   0.000    -.0656007   -.0420011
      select |   .1231439   .0428113     2.88   0.004     .0392353    .2070525
    articles |   .0509106   .0077581     6.56   0.000     .0357049    .0661163
     presthi |  -.9444709    .369606    -2.56   0.011    -1.668885   -.2200565
       _cons |  -5.770548   .3523052   -16.38   0.000    -6.461053   -5.080042
------------------------------------------------------------------------------
```

Example: gender differences in tenure

Odds ratios using `listcoef`

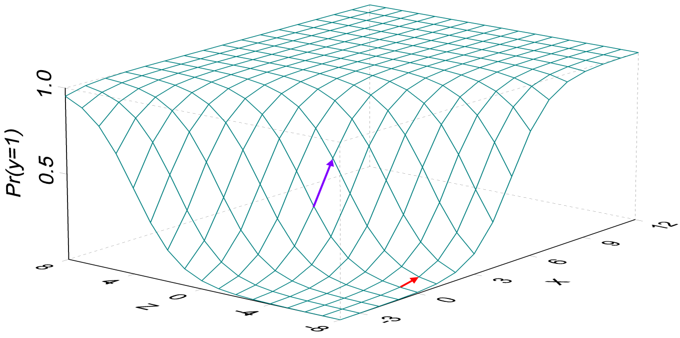
$$\text{Odds ratio for } \texttt{articles} = \exp\left(\beta_{\texttt{articles}}\right) = 1.05$$
$$\text{Odds ratio for } \texttt{female} = \exp\left(\beta_{\texttt{female}}\right) = 0.69$$

```
. listcoef articles female, help

logit (N=2945): Factor Change in Odds

  Odds of: Tenure vs NoTenure

----------------------------------------------------------------------
      tenure |      b         z     P>|z|    e^b    e^bStdX      SDofX
-------------+--------------------------------------------------------
    articles |   0.05091    6.562   0.000   1.0522   1.4097     6.7449
      female |  -0.37354   -2.941   0.003   0.6883   0.8341     0.4856
----------------------------------------------------------------------
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
     e^b = exp(b) = factor change in odds for unit increase in X
 e^bStdX = exp(b*SD of X) = change in odds for SD increase in X
   SDofX = standard deviation of X
```

Example: gender differences in tenure

Odds ratios compared to changes in probabilities

Both arrows represent the same factor change in the odds;
but the arrows represent very different changes in $\Pr(y = 1)$.



Predictions for a single set of x's

Use pseudo-observations for out of sample predictions with predict

$$\Pr\left(\texttt{tenure} = 1 \mid \mathbf{x}\right) = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots\right)}$$

```
 1. set obs 2946

    obs was 2945, now 2946

 2. replace female   =  1    if _n==2946
 3. replace articles =  0    if _n==2946
 4. replace year     =  7    if _n==2946
 5. replace yearsq   = 49    if _n==2946 // year squared
 6. replace select   =  4.97 if _n==2946 // mean level
 7. replace presthi  =  0.05 if _n==2946 // mean level

 8. predict prob in 2946

    (option p assumed; Pr(tenure))

 9. list prob female articles year yearsq select presthi in 2946

          prob    female    articles    year    yearsq    select    presthi
      .1559942         1           0       7        49      4.97        .05

10. drop in 2946

    (1 observation deleted)
```

We can make the same computation with `prvalue`.

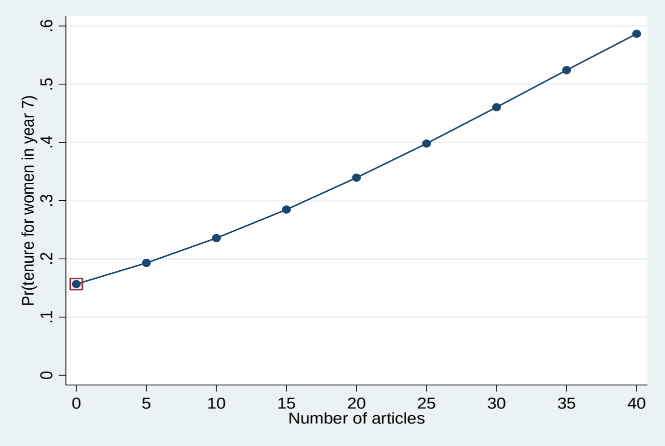
```
. prvalue, x(female=1 articles=0 year=7 yearsq=49) rest(mean)

logit: Predictions for tenure

Confidence intervals by delta method

                                  95% Conf. Interval
  Pr(y=Tenure|x):       0.1565   [ 0.1191,    0.1939]
  Pr(y=NoTenure|x):     0.8435   [ 0.8061,    0.8809]

         female        year      yearsq      select    articles     presthi
x=            1           7          49   4.9657725           0   .04516129
```

Next, we extend the use of predictions to demonstrate the "effect" of articles.

Plotting predictions

Probability of tenure for women in career year 7

The "effect" of `articles` at specific values of other variables:



Plotting predictions

Repeated calls to `prvalue` as the value of `articles` changes

Each point is computed by `prvalue`:

```
1. prvalue, x(articles=0 female=1 year=7 yearsq=49) brief

   logit: Predictions for tenure

                                   95% Conf. Interval
     Pr(y=Tenure|x):      0.1565   [ 0.1191,    0.1939]
     Pr(y=NoTenure|x):    0.8435   [ 0.8061,    0.8809]

2. prvalue, x(articles=5 female=1 year=7 yearsq=49) brief

   logit: Predictions for tenure

                                   95% Conf. Interval
     Pr(y=Tenure|x):      0.1931   [ 0.1552,    0.2311]
     Pr(y=NoTenure|x):    0.8069   [ 0.7689,    0.8448]

3. prvalue, x(articles=10 female=1 year=7 yearsq=49) brief

   logit: Predictions for tenure

                                   95% Conf. Interval
     Pr(y=Tenure|x):      0.2359   [ 0.1956,    0.2763]
     Pr(y=NoTenure|x):    0.7641   [ 0.7237,    0.8044]

4. prvalue, x(articles=15 female=1 year=7 yearsq=49) brief
   :::
```

Plotting predictions

Using results saved in r()

To automate things, we use information that `prvalue` saves in `r()`'s.

```
1. prvalue, x(articles= 0 female=1 year=7 yearsq=49) brief

   logit: Predictions for tenure

                                    95% Conf. Interval
     Pr(y=Tenure|x):      0.1565   [ 0.1191,    0.1939]
     Pr(y=NoTenure|x):    0.8435   [ 0.8061,    0.8809]

2. return list

   scalars:
               r(p1) =  .1565281003713608
            r(p1_lo) =  .1191238284620015
            r(p1_hi) =  .1939323722807201
            r(p0_lo) =  .8060676128181188
            :::

   matrices:
                r(x) :  1 x 6
```

Plotting predictions

Using forvalues with prvalue to collect predictions

Step 1: **Create variables** to hold the values to be plotted.

```
gen plotx = .
    label var plotx "Number of articles"
gen plotp1 = .
    label var plotp1 "Pr(tenure for women in year 7)"
```

Step 2: **Move predictions** from prvalue into these variables.

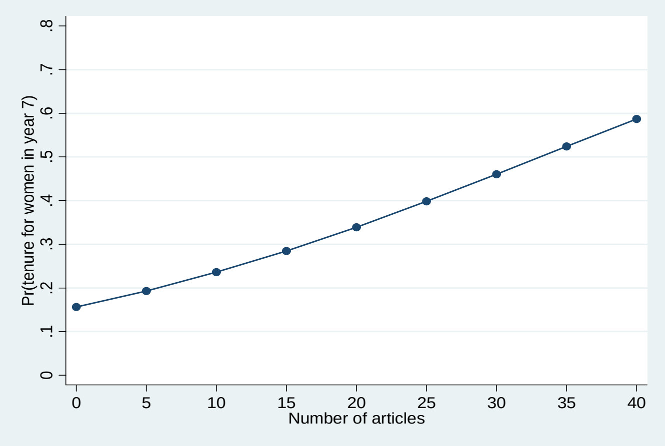
```
local i = 0
forvalues artval = 0(5)40 {
    local ++i
    quietly prvalue, x(articles=`artval' female=1 year=7 yearsq=49)
    replace plotx = `artval' if _n==`i'
    replace plotp1 = r(p1) if _n==`i'
}
```

Step 3: **Graph** the points:

```
graph twoway connected plotp1 plotx, ///
>   xlabel(0(5)40) ylabel(0(.1).8) ///
>   ytitle("Pr(tenure for women in year 7)")
```

Plotting predictions

Probability of tenure for women in career year 7



Confidence intervals for predicted probabilities

CIs can be computed using the delta method or the bootstrap

We can add confidence intervals to our predictions:

$$\left[\Pr\left(y = 1 \mid \mathbf{x}\right)_{\text{LowerBound}},\ \Pr\left(y = 1 \mid \mathbf{x}\right)_{\text{UpperBound}}\right]$$

1. **Delta method**: Computations are very quick using:

$$Var\left[\widehat{\Pr}\left(y = 1 \mid \mathbf{x}\right)\right] = \left[\frac{\partial F\left(\mathbf{x}\widehat{\boldsymbol{\beta}}\right)}{\partial \widehat{\boldsymbol{\beta}}}\right]^{T} Var(\widehat{\boldsymbol{\beta}}) \left[\frac{\partial F\left(\mathbf{x}\widehat{\boldsymbol{\beta}}\right)}{\partial \widehat{\boldsymbol{\beta}}}\right]$$

2. **Bootstrap method:** To get reliable results, you need to use at least 1,000 replications.

Step 1. **Generate predictions** using SPost's prgen:

```
1.    prgen articles, ci from(0) to(40) gap(5) generate(m0) ///
        x(female=1 year=7 yearsq=49)
```

Step 2. **Label variables** created by prgen:

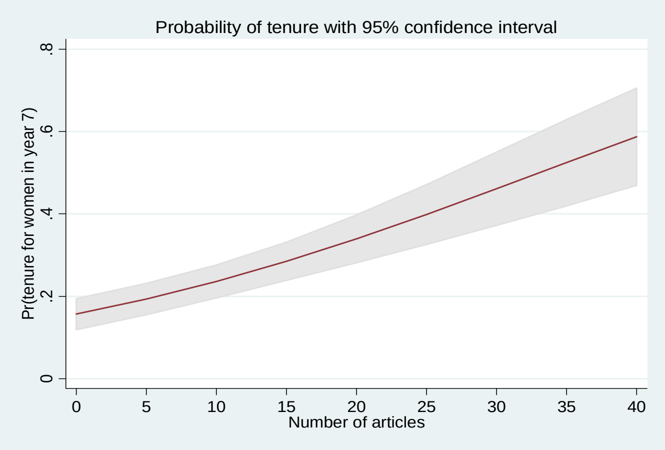
```
2.   label var m0p1 "Pr(tenure for women in year 7)"
3.   label var m0x  "Number of articles"
4.   label var m0p1lb "95% lower bound"
5.   label var m0p1ub "95% upper bound"
```

Step 3. **Plot** the results:

```
6.   graph twoway ///
 >      (rarea m0p1lb m0p1ub m0x, color(gs14)) ///
 >      (connected m0p1 m0x, msymbol(i)), ///
 >       subtitle("Probability of tenure with 95% confidence interval") ///
 >       yscale(range(0 .6)) ytitle("Pr(tenure for women in year 7)") ///
 >       xlabel(0(5)40) legend(off)
```

Confidence intervals for predicted probabilities

Plotting predictions and confidence intervals with SPost's `prgen`



Group comparisons - Overview

Methods for comparing groups

Approaches to make group comparisons.

1. **Include a dummy variable for group**: Include a dummy variable for the effect of group (e.g., $\beta_{\texttt{female}}$ in the prior model).
2. **Allow effects of x's to differ by group**: Allow the effects of x's to differ by group (e.g., let $\beta^{\text{men}}_{\texttt{articles}}$ and $\beta^{\text{women}}_{\texttt{articles}}$ differ).
3. **Test equality of coefficients?** Testing $\beta^{\text{men}}_{\texttt{articles}} = \beta^{\text{women}}_{\texttt{articles}}$ is problematic due to an identification problem.
4. **Compare predictions by across groups.**

In the LRM, we usually focus on comparing β's across groups.

1. For example,

$$\text{Men:} \quad y = \alpha^m + \beta^m_{educ} educ + \beta^m_{age} age + \varepsilon$$

$$\text{Women:} \quad y = \alpha^w + \beta^w_{educ} educ + \beta^w_{age} age + \varepsilon$$

2. Do men and women have the same return for education?

$$H_0\colon\ \beta^m_{educ} = \beta^w_{educ}$$

3. We compute:

$$z = \frac{\widehat{\beta}^m_{educ} - \widehat{\beta}^w_{educ}}{\sqrt{Var\left(\widehat{\beta}^m_{educ}\right) + Var\left(\widehat{\beta}^w_{educ}\right)}}$$

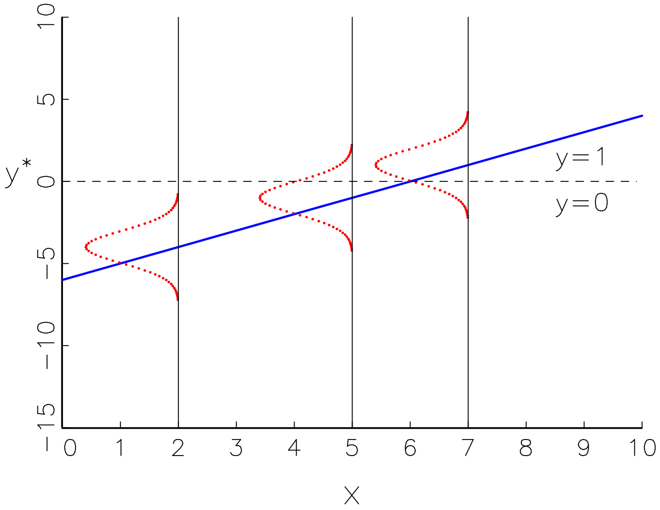
For binary and ordinal models, this approach does *not* work:

1. Since the β's and $Var(\varepsilon)$ are not seperately identifiable, the Chow test is inappropriate.
2. Alternatively, group comparisons of probabilities avoid this problem.
3. But, nonlinearity makes interpretations complicated.

The issue of identification is seen when the BRM is derived from an underlying latent variable y^*.

Logit and probit derived using a latent variable

Graphical representation



Logit and probit derived using a latent variable

Structural model predicting y*

1. **Structural model** with a latent y^*:

$$y^* = \alpha + \beta x + \varepsilon$$

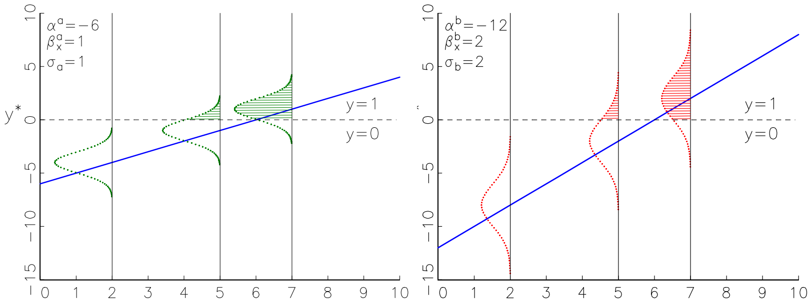
2. **Error** ε is normal(0,1) for probit; ε is logistic(0,$\pi^2/3$) for logit.
3. **y and y*** are linked by:

$$y = \begin{cases} 1 & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0 \end{cases}$$

4. **Pr(y=1)** depends on the error distribution *and* the coefficients:

$$\begin{aligned} \Pr(y = 1 \mid x) &= \Pr(y^* > 0 \mid x) \\ &= \Pr(\varepsilon < [\alpha + \beta x] \mid x) \end{aligned}$$

5. The identification problem can be seen graphically.



In terms of $\Pr(y=1)$, these are empirically indistinguishable:

Case 1: A change in x of 1 when $\beta_x^a = 1$ and $\sigma_a = 1$.

Case 2: A change in x of 1 when $\beta_x^b = 2$ and $\sigma_b = 2$.

1. Since y^* is not observed, β is only identified up to a scale factor.
2. Let y^* be the latent variable associated with receipt of tenure,

$$\begin{array}{ll} \text{Men:} & y^* = \alpha^m + \beta^m_{\texttt{articles}}\texttt{articles} + \varepsilon_m \\ \text{Women:} & y^* = \alpha^w + \beta^w_{\texttt{articles}}\texttt{articles} + \varepsilon_w \end{array}$$

3. Assume the "effects" of articles are equal:

$$\beta^m_{\texttt{articles}} = \beta^w_{\texttt{articles}}$$

4. And, assume women have more unobserved heterogeneity:

$$\sigma_w > \sigma_m$$

5. Now estimate the model...

Identification and group comparisons

Constraints imposed during estimation

1. Using `probit`, we assume that $\sigma = 1$.
2. For men, the estimated model for probit is:

$$
\begin{aligned}
\frac{y^*}{\sigma_m} &= \frac{\alpha^m}{\sigma_m} + \frac{\beta^m_{\texttt{articles}}}{\sigma_m}\texttt{articles} + \frac{\varepsilon_m}{\sigma_m} \\
&= \widetilde{\alpha}^m + \widetilde{\beta}^m_{\texttt{articles}}\texttt{articles} + \widetilde{\varepsilon}_m, \text{ where } \widetilde{\sigma}_m = 1
\end{aligned}
$$

3. For women, the estimated model for probit is:

$$
\begin{aligned}
\frac{y^*}{\sigma_w} &= \frac{\alpha^w}{\sigma_w} + \frac{\beta^w_{\texttt{articles}}}{\sigma_w}\texttt{articles} + \frac{\varepsilon_w}{\sigma_w} \\
&= \widetilde{\alpha}^w + \widetilde{\beta}^w_{\texttt{articles}}\texttt{articles} + \widetilde{\varepsilon}_w, \text{ where } \widetilde{\sigma}_w = 1
\end{aligned}
$$

4. Alternatively, `logit` assumes $\sigma = \pi/\sqrt{3}$.

1. Substantively, we want to test:

$$H_0\colon\ \beta^m_{\texttt{articles}} = \beta^w_{\texttt{articles}}\ .$$

2. But, we can only test:

$$H_0\colon\ \widetilde{\beta}^m_{\texttt{articles}} = \widetilde{\beta}^w_{\texttt{articles}}\ .$$

3. Unless the error variances are equal $(\sigma^2_m = \sigma^2_w)$,

$$\widetilde{\beta}^m_{\texttt{articles}} = \widetilde{\beta}^w_{\texttt{articles}}$$

does not imply

$$\beta^m_{\texttt{articles}} = \beta^w_{\texttt{articles}}\ .$$

Two distinct approaches address the identification problem.

1. Allison's (1999) test of H_0: $\beta_x^m = \beta_x^w$
   - Disentangles the β's and σ's.
   - But requires that $\beta_z^m = \beta_z^w$.
   - Rich Williams' `gologit2` implements this test.
2. Since the probabilities are invariant to σ, I propose testing

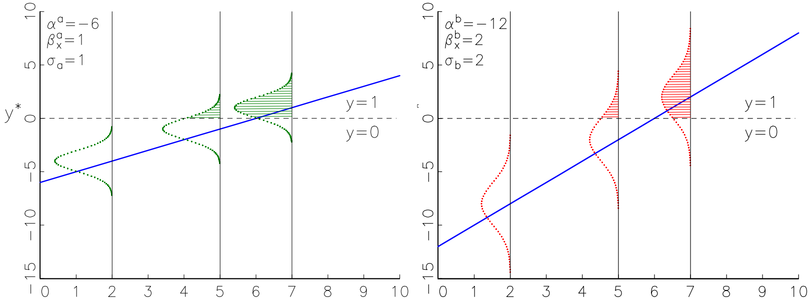
$$H_0\text{:}\ \Pr\left(y = 1 \mid \mathbf{x}^*\right)_m = \Pr\left(y = 1 \mid \mathbf{x}^*\right)_w$$

3. Graphically...

Comparing groups with logit and probit

Group comparisons of Pr(y=1)

Predicted probabilities are invariant to the assumed variance of ε:



Allow the effects of independent variables differ across groups:

1. Let $w = 1$ for women, else 0 and $wx = w \times x$;
   let $m = 1$ for men, else 0 and $mx = m \times x$.

$$\Pr(y = 1) = F\left(\alpha^{w} w + \beta_{x}^{w} wx + \alpha^{m} m + \beta_{x}^{m} mx\right)$$

2. Then:

$$\begin{aligned} \Pr(y = 1 \mid \mathbf{x})_{w} &= F\left(\alpha^{w} + \beta_{x}^{w} x\right) \\ \Pr(y = 1 \mid \mathbf{x})_{m} &= F\left(\alpha^{m} + \beta_{x}^{m} x\right) \end{aligned}$$

3. The gender difference in the probability of tenure is:

$$\Delta(\mathbf{x}) = \Pr(y = 1 \mid \mathbf{x})_{m} - \Pr(y = 1 \mid \mathbf{x})_{w}$$

M1: articles as the only predictor

Chow-type test confounds structural coefficients and unobserved heterogeneity

Start with a simple model with only publications predicting tenure:

```
 1. logit tenure female male f_articles m_articles, nolog nocon
    :::

 2. listcoef f_articles m_articles

logit (N=2945): Factor Change in Odds

  Odds of: Tenure vs NoTenure

----------------------------------------------------------------------
      tenure |      b         z     P>|z|     e^b    e^bStdX      SDofX
-------------+--------------------------------------------------------
  f_articles |    0.04215    4.259   0.000   1.0430    1.2855     5.9592
  m_articles |    0.09810    9.928   0.000   1.1031    1.7854     5.9089
----------------------------------------------------------------------

 3. test f_articles = m_articles // an incorrect test

     ( 1)  f_articles - m_articles = 0

              chi2(  1) =   16.01
            Prob > chi2 =    0.0001
```

M1: articles as the only predictor

Comparing predictions for men and women at a single value of articles

We can compute predictions along with differences using `prvalue`:

```
1. quietly prvalue, x(fem=1 f_art=5 male=0 m_art=0) save
2.         prvalue, x(fem=0 f_art=0 male=1 m_art=5) dif

   logit: Change in Predictions for tenure

   Confidence intervals by delta method

                         Current     Saved    Change   95% CI for Change
     Pr(y=Tenure|x):      0.0995    0.0943    0.0052  [-0.0179,   0.0284]
     Pr(y=NoTenure|x):    0.9005    0.9057   -0.0052  [-0.0284,   0.0179]

                  female  f_articles        male  m_articles
   Current=            0           0           1           5
     Saved=            1           5           0           0
      Diff=           -1          -5           1           5
```

Plotting predictions

Steps for creating graphs with predicted probabilities

Step 1. Compute predictions (`prvalue` to matrices).

$\Downarrow$

Step 2. Create variables with predictions (`svmat`).

$\Downarrow$

Step 3. Graph results (`graph`).

M1: articles as the only predictor

Saving probabilities to matrices and converting them to variables

Step 1. **Compute predictions** and put them in matrices.

```
foreach art of numlist 0(2)50 {

    quietly prvalue, x(fem=1 f_art=‘art’ male=0 m_art=0)
        matrix y_fem = nullmat(y_fem) \ pepred[2,2]
    quietly prvalue, x(fem=0 f_art=0 male=1 m_art=‘art’)
        matrix y_mal = nullmat(y_mal) \ pepred[2,2]
    matrix x_art = nullmat(x_art) \ ‘art’

}
```

Step 2. **Create variables** containing predictions.

```
svmat x_art
   label var x_art1 "Number of Articles"
svmat y_fem
   label var y_fem1 "Women" // Pr(for women)
svmat y_mal
   label var y_mal1 "Men"   // Pr(for men)
```

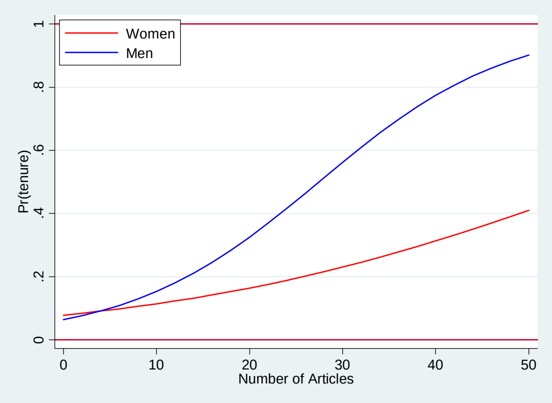
Step 3. **Graph** the results.

```
twoway (connected y_fem x_art, msym(i) clcol(red)) ///
>   (connected y_mal x_art, msym(i) clcol(blue))     ///
>  , ytitle(Pr(tenure)) xlabel(0(10)50)              ///
>    ylabel(0(.2)1.) yline(0 1) legend(pos(11) ring(0) cols(1))
```

M1: articles as the only predictor

Plotting predicted probabilities for men and women

This graph shows all of the predictions available from this model.



M1: articles as the only predictor

Comparing gender differences in predictions with confidence intervals

To compare groups at different levels of `articles`:

1. Compute differences:

$$\Delta\left(\texttt{articles}\right) = \Pr\left(y=1 \mid \texttt{articles}\right)_m - \Pr\left(y=1 \mid \texttt{articles}\right)_w$$

2. Confidence intervals are computed by delta or bootstrap:

$$\left[\Delta\left(\texttt{articles}\right)_{\text{LowerBound}}, \Delta\left(\texttt{articles}\right)_{\text{UpperBound}}\right]$$

3. With one RHS variable, we can plot all comparisons.

M1: articles as the only predictor

Computing confidence intervals for gender differences

Step 1. **Compute predictions** and save them in matrices:

```
 1.   foreach art of numlist 0(2)50 {

 2.       quietly prvalue, save                    /// for women
  >         x(fem=1 f_art=`art' male=0 m_art=0)
 3.       quietly prvalue, diff                    /// for men
  >         x(fem=0 f_art=0 male=1 m_art=`art')

 4.       matrix y_mal = nullmat(y_mal) \ pepred[2,2]
 5.       matrix y_fem = nullmat(y_fem) \ pepred[4,2]
 6.       matrix y_dc  = nullmat(y_dc)  \ pepred[6,2]
 7.       matrix y_ub  = nullmat(y_ub)  \ peupper[6,2]
 8.       matrix y_lb  = nullmat(y_lb)  \ pelower[6,2]
 9.       matrix x_art = nullmat(x_art) \ `art'

10.   }
```

M1: articles as the only predictor

Creating variables and setting up the graph

Step 2. **Create variables** with predictions for plotting:

```
foreach v in x_art y_dc y_ub y_lb y_fem y_mal {
    svmat `v'
}

label var x_art "Number of Articles"
label var y_fem "Women" // Pr(for women)
label var y_mal "Men"   // Pr(for men)
label var y_dc  "Difference" // Pr(for men) - Pr(for women)
label var y_ub  "95% confidence interval"
label var y_lb  "95% confidence interval"
```

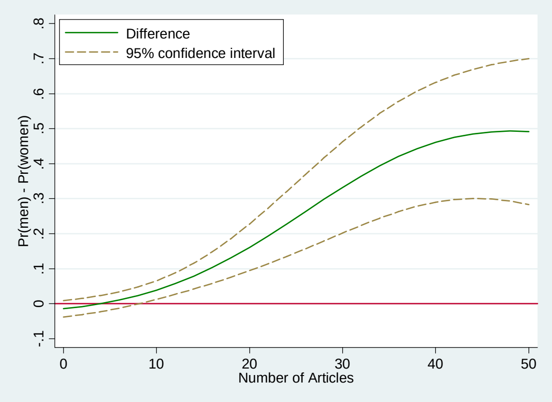
Step 3. **Graph** the results:

```
twoway ///
>   (connected y_dc x_art, msym(i) clcol(green))                 ///
>   (connected y_ub x_art, msym(i) clcol(brown) clpat(dash))     ///
>   (connected y_lb x_art, msym(i) clcol(brown) clpat(dash))     ///
>  , legend(pos(11) ring(0) cols(1) order(1 2)) xlabel(0(10)50) ///
>    ytitle("Pr(men) - Pr(women)") ylabel(-.1(.1).8) yline(0)
```

M1: articles as the only predictor

Plotting CIs for gender differences

This graph shows all of the predictions available from this model:



Adding variables introduces substantial complications:

1. With two independent variables:

$$\Pr(y = 1 \mid x, z) = F(\alpha + \beta_x x + \beta_z z)$$

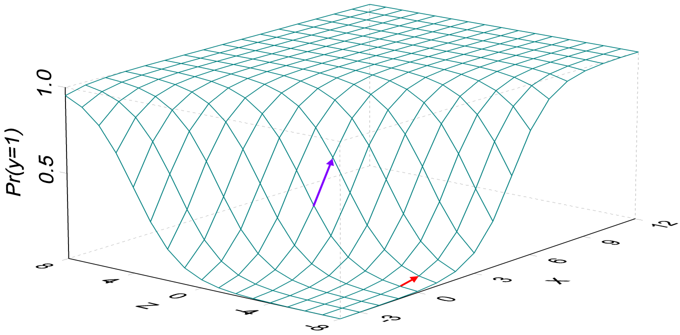
2. Setting $z = Z$ changes the intercept in an equation with only x:

$$\begin{aligned}\Pr(y = 1 \mid x, Z) &= F(\alpha + \beta_x x + \beta_z Z) \\ &= F([\alpha + \beta_z Z] + \beta_x x) \\ &= F(\alpha^* + \beta_x x)\end{aligned}$$

3. Predictions depend on the levels of each variable in the model.

Interpretation of regression models with Stata and SPost

Discrete changes depend on the level of other variables



Gender differences in the effect of x controlling for a single z:

1. For a given $z = Z$:

$$\begin{aligned} \text{Men:} \quad & \Pr(y = 1 \mid x, Z)_m = F(\alpha^{*m} + \beta_x^m x) \\ \text{Women:} \quad & \Pr(y = 1 \mid x, Z)_w = F(\alpha^{*w} + \beta_x^w x) \end{aligned}$$

2. Differences in probabilities for a given x depends on the level of other variables:

$$\Delta(x, Z) = \Pr(y = 1 \mid x, Z)_m - \Pr(y = 1 \mid x, Z)_w$$

Add a binary variable for a job in a high prestige department and estimate the model:

```
. logit tenure female f_art f_presthi ///
>                male m_art m_presthi, nolog nocon

Logistic regression                               Number of obs   =       2945
                                                  LR chi2(6)      =          .
Log likelihood = -1032.3002                       Prob > chi2     =          .

------------------------------------------------------------------------------
      tenure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.543769   .1421109   -17.90   0.000    -2.822302   -2.265237
  f_articles |   .0572428   .0114595     5.00   0.000     .0347826     .079703
   f_presthi |  -1.634833   .6719782    -2.43   0.015    -2.951886   -.3177794
        male |  -2.684516   .1173755   -22.87   0.000    -2.914568   -2.454464
  m_articles |   .1001105   .0099969    10.01   0.000     .0805168    .1197041
   m_presthi |  -.7295205   .4259048    -1.71   0.087    -1.564279    .1052376
------------------------------------------------------------------------------
```

M2: articles and having a prestigious job

Predictions for men & women at both levels of prestige

Step 1a. Compute gender differences for those with high prestige jobs:

```
 1.  foreach art of numlist 0(10)50 {

 2.      quietly prvalue, save   /// for women
  >        x(fem=1 f_art=`art' male=0 m_art=0     f_presthi=1 m_presthi=0)
 3.      quietly prvalue, diff   /// for men
  >        x(fem=0 f_art=0     male=1 m_art=`art' f_presthi=0 m_presthi=1)
 4.      matrix xlo_art = nullmat(xlo_art) \ `art'        // articles
 5.      matrix ylo_mal = nullmat(ylo_mal) \ pepred[2,2]  // pr men
 6.      matrix ylo_fem = nullmat(ylo_fem) \ pepred[4,2]  // pr women
 7.      matrix ylo_dc  = nullmat(ylo_dc)  \ pepred[6,2]  // difference
 8.      matrix ylo_ub  = nullmat(ylo_ub)  \ peupper[6,2] // upper limit
 9.      matrix ylo_lb  = nullmat(ylo_lb)  \ pelower[6,2] // lower limit

10.  }
```

Step 1b. Do the same thing for those not in high prestige jobs.

M2: articles and having a prestigious job

Plotting predicted probabilities at both levels of prestige for men and women

Step 2. Create variables with predictions:

```
1.  foreach v in xlo_art ylo_fem ylo_mal yhi_fem yhi_mal {
2.      svmat `v'
3.  }

4.  label var xlo_art1 "Number of Articles"
5.  label var ylo_fem1 "Women - not distinguished"
6.  label var ylo_mal1 "Men - not distinguished"
7.  label var yhi_fem1 "Women - distinguished"
8.  label var yhi_mal1 "Men - distinguished"
```

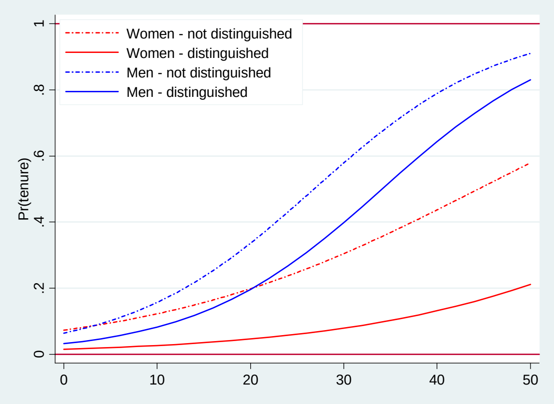
Step 3. Plot the results:

```
9.  twoway ///
 >    (con yhi_fem xhi_art, msym(i) clcol(red) clpat(solid)) ///
 >    (con ylo_fem xlo_art, msym(i) clcol(red) clpat(shortdash_dot)) ///
 >    (con yhi_mal xhi_art, msym(i) clcol(blue) clpat(solid)) ///
 >    (con ylo_mal xlo_art, msym(i) clcol(blue) clpat(shortdash_dot)) ///
 >    , legend(pos(11) order(2 1 4 3) ring(0) cols(1) region(ls(none))) ///
 >      ylabel(0(.2)1.) yline(0 1) ytitle(Pr(tenure)) xlabel(0(10)50)
```

M2: articles and having a prestigious job

Plotting predicted probabilities at both levels of prestige for men and women

This graph shows all of the predictions available from this model.



Alternatively, we can plot:

$$\Pr\left(\texttt{tenure} \mid \texttt{articles},\ \texttt{presthi}\right)_{\text{Men}} - \Pr\left(\texttt{tenure} \mid \texttt{articles},\ \texttt{presthi}\right)_{\text{Women}}$$

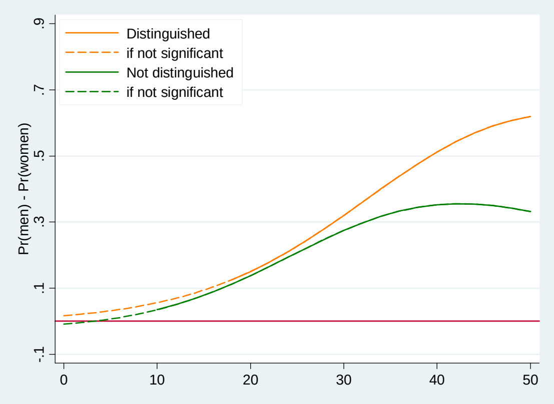
Non-significant differences are shown as dashed lines.

```
1.  gen ylo_sigdc = ylo_dc if ylo_lb>=0 & ylo_lb!=.
2.  gen yhi_sigdc = yhi_dc if yhi_lb>=0 & yhi_lb!=.
3.  label var ylo_sigdc "Not distinguished"
4.  label var ylo_dc    "if not significant"
5.  label var yhi_sigdc "Distinguished"
6.  label var yhi_dc    "if not significant"

7.  twoway ///
 >    (connected ylo_sigdc xlo_art, clpat(solid) msym(i) clcol(green) ) ///
 >    (connected yhi_sigdc xhi_art, clpat(solid) msym(i) clcol(orange)) ///
 >    (connected ylo_dc    xlo_art, clpat(dash)  msym(i) clcol(green) ) ///
 >    (connected yhi_dc    xhi_art, clpat(dash)  msym(i) clcol(orange)) ///
 >   , legend(pos(11) order(2 4 1 3) ring(0) cols(1) region(ls(none)))  ///
 >     ytitle("Pr(men) - Pr(women)") xlab(0(10)50) ylab(-.1(.2).9) ylin(0)
```

M2: articles and having a prestigious job

Plotting gender differences in probability of tenure at two levels of prestige

A graph provides all of the information from our model.



M3: articles, prestige, time in rank and other variables

Logit estimates using a continuous measure of prestige

Estimate a more complex model:

```
. logit tenure male m_year m_yearsq m_select m_articles m_prestige ///
>              fem  f_year f_yearsq f_select f_articles f_prestige, nolog nocon

Logistic regression                               Number of obs   =       2945
                                                  LR chi2(12)     =          .
Log likelihood = -918.07144                       Prob > chi2     =          .

      tenure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   -5.82375   .5041622   -11.55   0.000    -6.811889    -4.83561
      m_year |   1.071883   .1180005     9.08   0.000     .8406058    1.303159
    m_yearsq |  -.0654023   .0087056    -7.51   0.000     -.082465   -.0483397
    m_select |   .2107227   .0571501     3.69   0.000     .0987106    .3227349
  m_articles |   .0735537   .0107594     6.84   0.000     .0524656    .0946417
  m_prestige |  -.3770013    .103439    -3.64   0.000     -.579738   -.1742646
      female |  -4.207207    .630249    -6.68   0.000    -5.442473   -2.971942
      f_year |   .7685059   .1255128     6.12   0.000     .5225053    1.014507
    f_yearsq |  -.0417568   .0084699    -4.93   0.000    -.0583575   -.0251561
    f_select |   .0344378   .0683684     0.50   0.614    -.0995617    .1684373
  f_articles |   .0356986   .0119722     2.98   0.003     .0122335    .0591638
  f_prestige |  -.3481816    .152196    -2.29   0.022    -.6464803   -.0498829
```

M3: articles, prestige, time in rank and other variables

Mean level of predictors for holding other variables at group means

Compute gender differences as one variable changes, holding others constant.

Step 1a. Compute “constant” values with `summarize`:

```
foreach v in year yearsq select art prestige {
    quietly sum ‘v’
    local mn_‘v’ = r(mean)
}

local mn_yr = 7 // year for predictions
local mn_yrsq = ‘mn_yr’ * ‘mn_yr’ // year squared

local m_at_mn "m_year=‘mn_yr’ m_yearsq=‘mn_yrsq’ m_select=‘mn_select’"
local f_at_mn "f_year=‘mn_yr’ f_yearsq=‘mn_yrsq’ f_select=‘mn_select’"

local m_at_0  "mal=0 m_art=0 m_year=0 m_yearsq=0 m_select=0 m_prestige=0"
local f_at_0  "fem=0 f_art=0 f_year=0 f_yearsq=0 f_select=0 f_prestige=0"
```

Steps 1b & 2. With these control values, compute predictions and move results into variables:

M3: articles, prestige, time in rank and other variables

Discrete change in year 7 at 5 levels of prestige over range of articles

```
foreach p in 1 2 3 4 5 {  // loop over prestige
    foreach art of numlist 0(2)50 {  // loop over articles

        quietly prvalue, save ///
          x(fem=1 f_art=‘art’ f_prestige=‘p’ ‘f_at_mn’ ‘m_at_0’)
        quietly prvalue, diff ///
          x(mal=1 m_art=‘art’ m_prestige=‘p’ ‘m_at_mn’ ‘f_at_0’)

        matrix x‘p’_art = nullmat(x‘p’_art) \ ‘art’
        matrix y‘p’_mal = nullmat(y‘p’_mal) \ pepred[2,2]
        matrix y‘p’_fem = nullmat(y‘p’_fem) \ pepred[4,2]
        matrix y‘p’_dc  = nullmat(y‘p’_dc)  \ pepred[6,2]
        matrix y‘p’_ub  = nullmat(y‘p’_ub)  \ peupper[6,2]
        matrix y‘p’_lb  = nullmat(y‘p’_lb)  \ pelower[6,2]
    }

    foreach v in x‘p’_art y‘p’_dc y‘p’_ub y‘p’_lb y‘p’_fem y‘p’_mal {
        svmat ‘v’
    }
    label var x‘p’_art1 "Number of Articles"
    label var y‘p’_mal1 "Men"
    label var y‘p’_fem1 "Women"
    label var y‘p’_dc1  "Male-Female difference"
    label var y‘p’_ub1  "95% confidence interval"
    label var y‘p’_lb1  "95% confidence interval"

}
```

Step 3a. **Plot probabilities** for men and women:

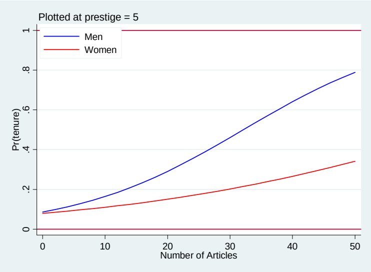
```
1.  twoway ///
 >    (connected y5_fem x5_art, msym(i) clcol(red)) ///
 >    (connected y5_mal x5_art, msym(i) clcol(blue)) ///
 >   , subtitle("Plotted at prestige = 5",pos(11)) ///
 >     legend(pos(11) order(2 1) ring(0) cols(1) region(ls(none))) ///
 >     ytitle(Pr(tenure)) xlabel(0(10)50) ///
 >     ylabel(0(.2)1.) yline(0 1)
```

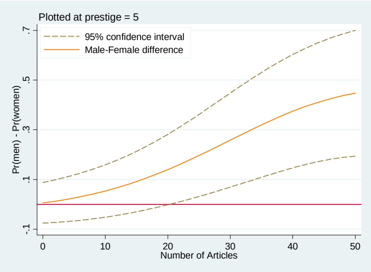
Step 3b. Or, **plot differences** in probabilities for men and women:

```
2.  twoway ///
 >    (connected y5_dc x5_art, msym(i) clcol(orange)) ///
 >    (connected y5_ub x5_art, msym(i) clcol(brown) clpat(dash)) ///
 >    (connected y5_lb x5_art, msym(i) clcol(brown) clpat(dash)) ///
 >   , subtitle("Plotted at prestige = 5",pos(11)) ///
 >     legend(pos(11) order(2 1) ring(0) cols(1) region(ls(none))) ///
 >     ytitle("Pr(men) - Pr(women)") xlabel(0(10)50) ///
 >     ylabel(-.1(.2).7) yline(0)
```

M3: articles, prestige, time in rank and other variables

Plot of probability and discrete change in year 7 with prestige 5





M3: articles, prestige, time in rank and other variables

Discrete change in year 7 at five prestige levels

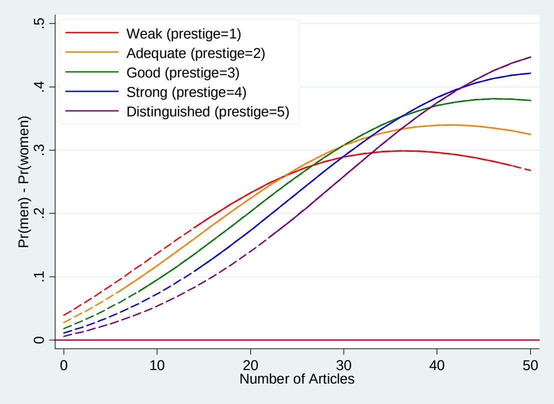
Step 3c. Let dashed lines indicate non-significant differences and plot five levels of prestige in the same graph:

```
 1.  foreach p in 1 2 3 4 5 {
 2.      gen y`p'_sigdc = y`p'_dc if y`p'_lb>=0 & y`p'_lb!=.
 3.  }

 4.  label var y1_sigdc "Weak (prestige=1)"
     ... and so on ...

12. twoway ///
  >   (connected y1_sigdc x1_art, clpat(solid) msym(i) clcol(red))    ///
  >   (connected y1_dc    x1_art, clpat(dash)  msym(i) clcol(red))    ///
  >   (connected y2_sigdc x2_art, clpat(solid) msym(i) clcol(orange)) ///
  >   (connected y2_dc    x2_art, clpat(dash)  msym(i) clcol(orange)) ///
  >   (connected y3_sigdc x3_art, clpat(solid) msym(i) clcol(green))  ///
  >   (connected y3_dc    x3_art, clpat(dash)  msym(i) clcol(green))  ///
  >   (connected y4_sigdc x4_art, clpat(solid) msym(i) clcol(blue))   ///
  >   (connected y4_dc    x4_art, clpat(dash)  msym(i) clcol(blue))   ///
  >   (connected y5_sigdc x5_art, clpat(solid) msym(i) clcol(purple)) ///
  >   (connected y5_dc    x5_art, clpat(dash)  msym(i) clcol(purple)) ///
  >  , legend(pos(11) order(1 2 3 4 5) ring(0) cols(1) region(ls(none))) ///
  >    ytitle("Pr(men) - Pr(women)") xlab(0(10)50) ylab(.0(.1).5) ylin(0)
```

M3: articles, prestige, time in rank and other variables

Discrete change in year 7 at five prestige levels as number of articles varies

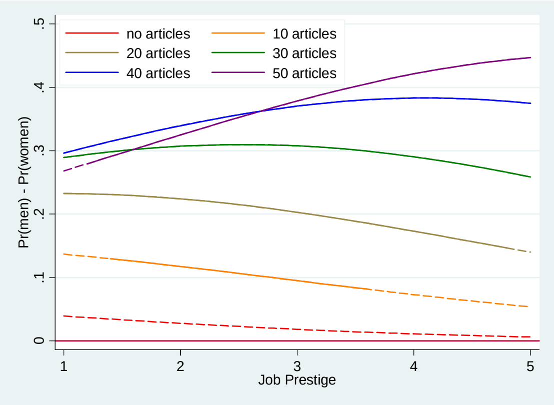
Holding all other variables constant, we can assess the effects of three variables on tenure:



M3: articles, prestige, time in rank and other variables

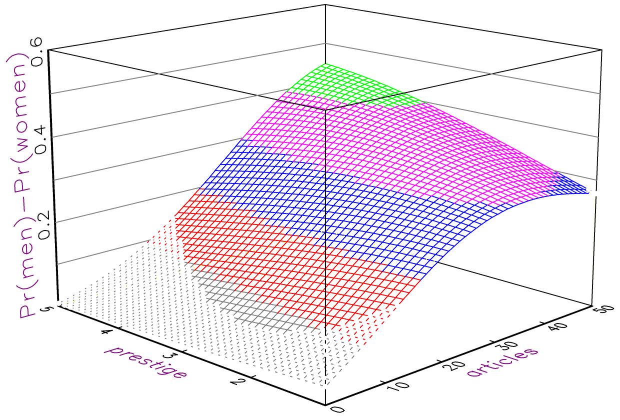
Discrete change in year 7 at six levels of article as prestige varies

The same (!) information can be presented by reversing the way we use job prestige and articles in the graph:



M3: articles, prestige, time in rank and other variables

Discrete change for prestige and number of articles in year 7



Installing the SPost programs

Using SPost

```
. net from http://www.indiana.edu/~jslsoc/stata/

------------------------------------------------------------------------
:::
http://www.indiana.edu/~jslsoc/stata/
SPost: post-estimation interpretation of regression models.
:::

:::
PACKAGES you could -net describe-:
 spost9_ado : Stata 9 SPost ado files.
 spost9_do : Stata 9 SPost sample do and dta files.
 spost_groups : Long and Xu - comparing group differences.
:::

------------------------------------------------------------------------

. findit spost9
```

References

1. Allison, Paul D. 1999. "Comparing Logit and Probit Coefficients Across Groups." Sociological Methods and Research 28:186-208.
2. Chow, G.C. 1960. "Tests of equality between sets of coefficients in two linear regressions." Econometrica 28:591-605.
3. Long, J.S. and Freese, J. 2005. Regression Models for Categorical and Limited Dependent Variables with Stata. Second Edition. College Station, TX: Stata Press.
4. Long, J. Scott, Paul D. Allison, and Robert McGinnis. 1993. "Rank Advancement in Academic Careers: Sex Differences and the Effects of Productivity." American Sociological Review 58:703-722.
5. Xu, J. and J.S. Long, 2005, Confidence intervals for predicted outcomes in regression models for categorical outcomes. The Stata Journal 5: 537-559.