Group comparisons in logit and probit using predicted probabilities[1]

J. Scott Long
Indiana University

June 25, 2009

Abstract

The comparison of groups in regression models for binary outcomes is complicated by an identification problem inherent in these models. Traditional tests of the equality of coefficients across groups confound the magnitude of the regression coefficients with residual variation. If the amount of residual variation differs between groups, the test can lead to incorrect conclusions (Allison 1999). Allison proposes a test for the equality of regression coefficients that removes the effect of group differences in residual variation by adding the assumption that the regression coefficients for some variables are identical across groups. In practice, a researcher is unlikely to have either empirical or theoretical justification for this assumption, in which case the Allison's test can also lead to incorrect conclusions. An alternative approach, suggested here, 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. Using predicted probabilities requires researchers to think differently about comparing groups. With tests of the equality of regression coefficients, a single test lets the researcher conclude easily whether the effects of a variable are equal across groups. Testing the equality of predicted probabilities requires multiple tests since group differences in predictions vary with the levels of the variables in the model. A researcher must examine group differences in predictions at multiple levels of the variables often requiring more complex conclusions on how groups differ in the effect of a variable.

[1] I thank Paul Allison, Ken Bollen, Rafe Stolzenberg, Pravin Trivedi, and Rich Williams for their comments.

Group comparisons in logit and probit using predicted probabilities

1 Overview

The comparison of groups is fundamental to research in many areas and tests comparing groups have received a great deal of attention. Chow's (1960) paper, declared a "citation classic" (Garfield 1984), provides a general framework for group comparisons in the linear regression model. Suppose that we are comparing the effect of x on y for women and men, where β_x^W and β_x^M are the coefficients of interest. If H_0: $\beta_x^W = \beta_x^M$ is rejected, we conclude that the effect of x differs for men and women. This approach to testing group differences can be applied to many types of regression models as shown by Liao (2002). Allison (1999) points out a critical problem when this test is used in models such as logit or probit. For these models, standard tests can lead to incorrect conclusions since they confound the magnitude of the regression coefficients with the amount of residual variation. Allison proposes a test that removes the effect of residual variation by assuming that the coefficients for at least one independent variable are the same in both groups. Unfortunately, a researcher might lack sufficient theoretical or empirical information to justify such an assumption. Making an *ad hoc* decision that some regression coefficients are equal can lead to incorrect conclusions.[2] Tests of predicted probabilities provide an alternative approach for comparing groups that is unaffected by group differences in residual variation and does not require assumptions about the equality of regression coefficients for some variables. Group comparisons are made by testing the equality of predicted probabilities at different values of the independent variables.

This paper begins by reviewing why standard tests of the equality of regression coefficients across groups are inappropriate in some types of models. I then show why predicted probabilities are unaffected by residual variation and present a test of the equality of predicted probabilities that can be used for group comparisons of effects in models such as logit and probit. To illustrate this approach, I begin with a model that includes a single independent variable, in which case all information about group differences can be shown in a simple graph. I extend this approach to models with multiple independent variables, which requires more complex analysis due to the nonlinearity of the models.

[2] Williams (2009) raises other concerns with this tests that are discussed below.

2 Model identification in logit and probit

Discussions of the logit and probit model often note that the slope coefficients are only identified up to a scale factor (Maddala 1983:23). This lack of identification is why standard tests of the equality of regression coefficients across groups should not be used. To see how identification causes a problem in group comparisons and why predicted probabilities are not affected by this problem, consider how these models are derived using an underlying latent variable (see Long 1997: 40-50 for a full derivation). Suppose that the latent y^* is linearly related to an observed x through the structural model

$$y^* = \alpha_0 + \alpha_1 x + \varepsilon,$$

where I use a single independent variable for simplicity. The latent y^* is linked to an observed, binary y by the measurement equation

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

If y^* is greater than 0, y is observed as 1. Otherwise, y is observed as 0. For example, when a person's propensity to be in the labor force exceeds 0, she joins the labor force (i.e., $y = 1$). If her propensity is at or below 0, she is not in the labor force (i.e., $y = 0$).

According to equation 1, the probability of $y = 1$ is the proportion of the distribution of y^* that is above 0 at a given value of x:

$$\Pr\left(y = 1 \mid x\right) = \Pr\left(y^* > 0 \mid x\right). \tag{2}$$

Substituting $y^* = \alpha_0 + \alpha_1 x + \varepsilon$ and rearranging terms, the probability can be written in terms of the distribution of the errors:

$$\Pr(y = 1 \mid x) = \Pr(\varepsilon \leq \alpha_0 + \alpha_1 x \mid x).$$

[Figures 1 & 2 about here]

Figure 1 illustrates the relationship between the structural model for y^* and the probability of y for specific values of the parameters labeled as set A:

$$\begin{aligned} y_A^* &= \alpha_0^A + \alpha_1^A x + \varepsilon_A \\ &= -6 + 1x + \varepsilon_A \ . \end{aligned} \tag{3}$$

Assume that ε_A is normally distributed with mean 0 and variance $\sigma_A^2 = 1$, which are the usual assumptions for the probit model. The probability of $y = 1$ at a given value of x is the shaded portion of the error distribution in Figure 1 corresponding to

$$\Pr\left(y = 1 \mid x\right) = \Pr(\varepsilon_A \leq \alpha_0^A + \alpha_1^A x \mid x). \tag{4}$$

If y_A^* was observed, as is the case for the dependent varaible in the linear regression model, we could estimate the variance of ε_A. But, since both ε_A and y_A^* are unobserved, the variance of ε_A is unidentified and must be assumed. The lack of identification can be shown by changing the parameters in equation 3 by any nonzero factor δ:

$$\delta y_A^* = \delta\alpha_0^A + \delta\alpha_1^A x + \delta\varepsilon_A.$$

For example, suppose that $\delta = 2$. Then

$$2y_A^* = 2\alpha_0^A + 2\alpha_1^A x + 2\varepsilon_A.$$

Defining $y_B^* \equiv 2y_A^*$, $\alpha_0^B \equiv 2\alpha_0^A$, $\alpha_1^B \equiv 2\alpha_1^A$, and $\varepsilon_B \equiv 2\varepsilon_A$, the structural equation can be written as

$$\begin{aligned} y_B^* &= \alpha_0^B + \alpha_1^B x + \varepsilon_B \\ &= -12 + 2x + \varepsilon_B \end{aligned} \tag{5}$$

where $\sigma_B = 2\sigma_A = 2$. While the slope, intercept, and standard deviation of the errors in equations 3 and 5 differ by a factor of $\delta = 2$, the predicted probabilities are unaffected since changing the parameters by a scale factor simply "stretches" Figure 1, as shown in Figure 2. With set B of parameters, we are multiplying the inequality in equation 4 by $\delta = 2$, which does not change the probability of $y = 1$:

$$\begin{aligned} \Pr(y = 1 \mid x) &= \Pr(2\varepsilon_A \leq 2\alpha_0^A + 2\alpha_1^A x \mid x) \\ &= \Pr(\varepsilon_B \leq \alpha_0^B + \alpha_1^B x \mid x)\ . \end{aligned}$$

While the magnitude of the α's and σ are affected by the factor change, the predicted probabilities are not. Even if the magnitude of the α's and σ are of substantive interest, they cannot be estimated individually since all scale factors of the parameters have the same predictions for the observed data. That is, the α's are only identified up to a scale factor. Typically, to identify the model, the variances of the errors are assumed to have a single, specific value.

This is the fundamental issue behind Allison's critique of standard tests to compare coefficients across groups in logit and probit models. To see this, consider a pair of models for men and women:

$$\text{Men:} \qquad y^* = \alpha_0^M + \alpha_1^M x + \varepsilon_M \text{ where } \varepsilon_M \sim \phi\left(0, \sigma_M^2\right) \tag{6}$$

$$\text{Women:} \qquad y^* = \alpha_0^W + \alpha_1^W x + \varepsilon_W \text{ where } \varepsilon_W \sim \phi\left(0, \sigma_W^2\right) \tag{7}$$

Suppose that the slopes for x are identical, but that residual variation is greater for women than men. That is, $\alpha_1^M = \alpha_1^W$ and $\sigma_W^2 > \sigma_M^2$. If y^* was *observed*, this would be a pair of linear regressions with identical slopes for x. Since the variance of the errors is larger for women than men, the coefficient of determination R^2 would be smaller for women. But, since y^* is latent, the variance of errors must be assumed. While the variance could be fixed to any value, the usual (but arbitrary) assumption for probit is that $\sigma^2 = 1$ and for logit that $\sigma^2 = \pi^2/3$. Here, for simplicity and without loss of generality, I assume that $\sigma^2 = 1$. Critically, the same identifying assumption is used for both men and women. That is, both men and women are assumed to have the save variance for ε and accordingly the same R^2. To apply this assumption to equation 6 (i.e., to force the variance of ε to be 1), I divide the equation by σ_M. For men,

$$\frac{y^*}{\sigma_M} = \frac{\alpha_0^M}{\sigma_M} + \frac{\alpha_1^M}{\sigma_M}x + \frac{\varepsilon_M}{\sigma_M} \ .$$

To simplify notation, I use β's for the rescaled parameters and τ for the rescaled error:

$$\frac{y^*}{\sigma_M} = \beta_0^M + \beta_1^M x + \tau_M \tag{8}$$

where the variance of τ_M is 1. Similarly for women, I transform equation 7:

$$\begin{aligned} \frac{y^*}{\sigma_W} &= \frac{\alpha_0^W}{\sigma_W} + \frac{\alpha_1^W}{\sigma_W}x + \frac{\varepsilon_W}{\sigma_W} \\ &= \beta_0^W + \beta_1^W x + \tau_W \end{aligned} \tag{9}$$

where τ_W also has a unit variance by construction.

Software for probit as found in standard packages such as Stata, SPSS and SAS incorporates the identifying assumption that the variance of the errors is 1 and estimates equations 8 and 9, *not* equations 6 and 7.[3] That is, the software estimates the rescaled β-parameters, not the original α-parameters. A standard test of the equality of regression coefficients across groups evaluates the hypothesis

$$H_A\colon\ \frac{\alpha_1^M}{\sigma_M} = \frac{\alpha_1^W}{\sigma_W}$$

or equivalently

$$H_A\colon\ \beta_1^M = \beta_1^W,$$

not the hypothesis

$$H_B\colon\ \alpha_1^M = \alpha_1^W.$$

Moreover, H_A and H_B are only equivalent if $\sigma_W = \sigma_M$. This is why standard tests confound the slopes and the variances of the errors which can lead to incorrect conclusions.

To test the cross group equality of the effect of x without confounding group differences in residual variation, that is, to test H_B: $\alpha_1^M = \alpha_1^W$, requires information

[3] Standard software for logit rescales the errors to $\tau = \frac{\pi\varepsilon}{\sqrt{3}\sigma}$ so that $\text{Var}(\tau) = \pi^2/3$.

about group differences in the variance of the errors. Allison's test obtains this information by *assuming* that the coefficients for at least one variable are the same for both groups. For example, suppose that our model also include the predictor z:

$$\begin{aligned} \text{Men:} \quad & y^* = \alpha_0^M + \alpha_1^M x + \alpha_2^M z + \varepsilon_M \text{ where } \varepsilon_M \sim \phi\left(0, \sigma_M^2\right) \\ \text{Women:} \quad & y^* = \alpha_0^W + \alpha_1^W x + \alpha_2^W z + \varepsilon_W \text{ where } \varepsilon_W \sim \phi\left(0, \sigma_W^2\right). \end{aligned}$$

If the slope coefficient for z is the same for both groups (i.e., $\alpha_2^M = \alpha_2^W = \alpha_2$), then the ratio of the rescaled β-coefficients for z equals the ratio of the standard deviations of the errors:

$$\frac{\beta_2^M}{\beta_2^W} = \frac{\left(\alpha_2^M/\sigma_M\right)}{\left(\alpha_2^W/\sigma_W\right)} = \frac{\left(\alpha_2/\sigma_M\right)}{\left(\alpha_2/\sigma_W\right)} = \frac{\sigma_W}{\sigma_M}.$$

This provides the information needed to test H_B: $\alpha_1^M = \alpha_1^W$.

[Figure 3 about here]

There are two issues to keep in mind when using Allison's test. First, the test depends critically on the assumption that the effects of at least one variable are equal across groups. If the equality assumption is justified, Allison's test allows you to test the equality of the α-coefficients across groups in logit and probit without contamination by group differences in residual variation. In practice, however, I think it is unlikely that researchers will have sufficient information to justify the assumption that the effects of some variables are zero. If the equality assumption is unjustified, Allison's test can lead to incorrect conclusions. Using the same basic approach, Williams (2009) extends and refines the tests proposed by Allison by showing how they can be incorporated into the heterogeneous choice model.

A second issue related to Allison's approach is that the equality of regression coefficients across groups has different implications in nonlinear models than in linear models. In a linear model, if the coefficient for x is the same for men and women, both men and women receive the same expected change in y for a given change in x holding all other variables constant. For example, for each additional year of education both men and women are expected to increase their wages by β_x dollars, holding all else constant.[4] In a nonlinear model, having the same β_x for two groups does not imply that a change in x will have the same change in $\Pr(y = 1)$ for both groups. This is shown in Figure 3 which plots a probit or logit model where men and women have the same coefficient for x while the intercepts differ. As x increases from 2 to 3, the change in predicted probability for men, Δ_{Men}, is much larger than the change Δ_{Women} for women. In nonlinear models, having equal regression parameters for a given variable does not imply similar changes in the predicted probabilities.

[4] Significantly, in the linear regression model the equality of regression coefficients for two groups does not imply that the R^2's for two groups are the same.

3 Testing group differences in predicted probabilities

While regression coefficients are affected by the identifying assumption for the variance of the errors, the predicted probabilities are not. Accordingly, if you estimate logit and probit models using the same data, the estimated β's for the logit model will be about $\pi/\sqrt{3}$ times larger than those from the probit model because the assumed standard deviation of the errors for logit is $\pi/\sqrt{3}$ times larger than the assumed standard deviation for probit. The predicted probabilities, however, will be nearly identical. They differ only because the shape of the normal distribution used in probit is slightly different than the shape of the logistic distribution assumed for logit. Since predicted probabilities are not affected by group differences in residual variation, you can compare groups by testing the equality of predicted probabilities at substantively interesting values of the independent variables.

Tests of predicted probabilities can be computed with the delta method (Xu and Long 2005). The delta method computes the variance of functions of maximum likelihood estimates by creating a linear approximation of a function and computing the variance of the simpler, linear function (Cameron and Trivedi 2005:227-233). Let $\widehat{\boldsymbol{\beta}}$ be a vector of maximum likelihood estimates and let $G\left(.\right)$ be a function of $\boldsymbol{\beta}$, such as predicted probabilities. Using a Taylor series expansion, $G(\widehat{\boldsymbol{\beta}}) \approx G(\boldsymbol{\beta}) + \frac{\partial G(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Under standard assumptions, $G(\widehat{\boldsymbol{\beta}})$ is distributed normally around $G(\boldsymbol{\beta})$ with a variance of

$$Var\left[G(\widehat{\boldsymbol{\beta}})\right] = \frac{\partial G(\widehat{\boldsymbol{\beta}})}{\partial \widehat{\boldsymbol{\beta}}'} Var(\widehat{\boldsymbol{\beta}}) \frac{\partial G(\widehat{\boldsymbol{\beta}})}{\partial \widehat{\boldsymbol{\beta}}} \ , \tag{10}$$

where $Var(\widehat{\boldsymbol{\beta}})$ is the covariance matrix for the estimated parameters.

To apply the delta method to predicted probabilities, it helps to use the shorthand notation that $\pi\left(\mathbf{x}_i\right) \equiv \Pr\left(y_i = 1 \mid \mathbf{x}_i\right)$. If $G\left(\boldsymbol{\beta}\right) = \pi\left(\mathbf{x}_i\right) = F(\mathbf{x}_i'\boldsymbol{\beta})$ where $F\left(.\right)$ is the cdf for the standard normal distribution for the probit model and the standardized logistic distribution for the logit model. Then

$$\begin{aligned} \frac{\partial G(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial F(\mathbf{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= f\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\mathbf{x}_i \end{aligned}$$

where $f\left(.\right)$ is the pdf. Then,

$$\begin{aligned} Var\left[\pi\left(\mathbf{x}_i\right)\right] &= \frac{\partial F(\mathbf{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} Var(\widehat{\boldsymbol{\beta}}) \frac{\partial F(\mathbf{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= f\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\mathbf{x}_i' Var(\widehat{\boldsymbol{\beta}})\mathbf{x}_i f\left(\mathbf{x}_i'\boldsymbol{\beta}\right). \end{aligned}$$

The variance of the group differences in probabilities is

$$Var\left[\pi\left(\mathbf{x}_1^*\right)^{\text{Group 1}} - \pi\left(\mathbf{x}_2^*\right)^{\text{Group 2}}\right] = Var\left[\pi\left(\mathbf{x}_1^*\right)^{\text{Group 1}}\right] + Var\left[\pi\left(\mathbf{x}_2^*\right)^{\text{Group 2}}\right],$$

where $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ indicate values of the x's for the two groups. These values could be identical (e.g., test the predicted probabilities for the same levels of all variables) or different (e.g., test the predicted probabilities when the values of the x's for group 1 are the means for group 2 and vice versa). The z-statistic to test H_0: $\pi\left(\mathbf{x}_1^*\right)^{\text{Group 1}} = \pi\left(\mathbf{x}_2^*\right)^{\text{Group 2}}$ is

$$z = \frac{\pi\left(\mathbf{x}_1^*\right)^{\text{Group 1}} - \pi\left(\mathbf{x}_2^*\right)^{\text{Group 2}}}{\sqrt{Var\left[\pi\left(\mathbf{x}_1^*\right)^{\text{Group 1}} - \pi\left(\mathbf{x}_2^*\right)^{\text{Group 2}}\right]}}\ , \tag{11}$$

which has an asymptotic normal distribution. Defining $\Delta\left(\mathbf{x}\right) \equiv \pi\left(\mathbf{x}_1^*\right)^{\text{Group 1}} - \pi\left(\mathbf{x}_2^*\right)^{\text{Group 2}}$, a 95% confidence interval can be constructed such that

$$\Pr\left(\Delta\left(\mathbf{x}\right)_{\text{Lower Bound}} \leq \Delta\left(\mathbf{x}\right) \leq \Delta\left(\mathbf{x}\right)_{\text{Upper Bound}}\right) = .95\ .$$

Further details on the derivation are provided in the appendix.

4 Example

The following example is based on 301 male and 177 female biochemists who obtained Ph.D.s in the late 1950s and early 1960s and who held faculty positions in graduate departments (see Long, Allison, and McGinnis (1993) for further details). Following Allison (1999), the unit of analysis is person-years with each observation corresponding to one year in rank for one scientist with multiple records for each scientist (one record for each year a scientist is an assistant professor). The dependent variable is coded 1 if a scientist obtained tenure in the given year, else 0. Independent variables include the number of years a scientist has been in the rank of assistant professor, the selectivity of the scientist's undergraduate institution, the cumulative number of articles published by the end of each person-year, and the prestige of the department in which the scientist is working. Descriptive statistics are given in Table 1. Using these data, three increasingly complex models are estimated and groups differences examined. Model 1 uses only gender and the number of articles to predict tenure. Model 2 adds a binary indictor of having a job in a distinguished department. Model 3 includes year, year-squared, selectivity, total articles, and departmental prestige. Analyses were completed using Stata Version 10 (2007) along with Long and Freese's (2005) SPost commands.[5]

[5] The GC (group comparison) package by Long(2009) is a wrapper for the SPost command `prvalue` that makes estimating group comparisons much simpler.

[Tables 1 & 7 and Figures 4 & 5 about here]

Model 1 uses only the number of articles to predict tenure:

$$\begin{aligned}\pi^{M}\left(\mathit{articles}\right) &= \Lambda(\beta_0^M + \beta_1^M \mathit{articles})\\ \pi^{W}\left(\mathit{articles}\right) &= \Lambda(\beta_0^W + \beta_1^W \mathit{articles})\end{aligned}$$

with estimated β's given in Table 2. Based on these estimates, the predicted probability of tenure for men with 15 articles is .23 and for women is .14, a difference of .09. To test if the difference is significant, I compute a z-test using equation 11:

$$\begin{aligned} z &= \frac{\widehat{\pi}^{M}\left(\mathit{articles{=}15}\right) - \widehat{\pi}^{W}\left(\mathit{articles{=}15}\right)}{\sqrt{\mathit{Var}\left[\widehat{\pi}^{M}\left(\mathit{articles{=}15}\right) - \widehat{\pi}^{W}\left(\mathit{articles{=}15}\right)\right]}}\\ &= 4.11\end{aligned}$$

I reject the null hypothesis H_0: $\pi^{M}\left(\mathit{articles{=}15}\right) = \pi^{W}\left(\mathit{articles{=}15}\right)$ at the .01 level. Since the logit model is nonlinear, gender differences in the predicted probability of tenure vary by the number of articles. This is shown in Figure 4 which plots the probability of tenure for men and women with articles ranging from 0 to 50. Differences in probability are small for scientists with fewer publications. For example, for scientists with 7 publications, the difference is only .016 which is not significantly different from 0 ($z = 1.32$). To examine gender differences throughout the range of articles, I can plot $\widehat{\pi}^{M}\left(\mathit{articles}\right) - \widehat{\pi}^{W}\left(\mathit{articles}\right)$ along with the 95% confidence interval as shown in Figure 5. If the confidence interval extends below 0, the difference is not significant. Overall, gender differences in the probability of tenure are not significant for those with less than 10 publications and get steadily larger until about 40 publications at which point they level off with a difference of about .45.

This simple example illustrates a critical difference between tests of regression coefficients and tests using predicted probabilities. With tests of regression coefficients, there is a single test of whether the effect of a variable differs across groups.[6] With predicted probabilities, there are multiple tests for different levels of the independent variables. As we just saw, the group differences can be significant at some values of the predictor, while not significant at others. This complicates things for the data analyst since there might not be a simple answer to the question of whether the effects of a variable are the same for both groups. But there are also advantages. First, using probabilities avoids assumptions about the equality of the effects of other variables. Second, the more complex answer to whether the effects of a variable differ across groups could be a more realistic representation of the phenomenon being modeled. In

[6] Note that Allison's test can not be used with a single predictor since there is no other variable in the model whose effect can be assumed equal across groups.

this example, it makes substantive sense that gender differences are small and nonsignificant at lower levels of productivity since neither men nor women are likely to be tenured if they have not been productive, but there could gender discrimination at higher levels of productivity where tenure is more likely.

[Table 3 and Figures 6, 7 & 8 about here]

With a single independent variable, a simple graph shows all of the differences between men and women in the effect of articles on tenure. With additional variables, interpretation is more complicated since group differences in predictions differ by the level of all variables in the model. This is a critical point for understanding how probabilities are used to compare groups. To show this, Model 2 retains the number of articles and adds a binary variable indicating whether a scientist is located in a distinguished department, leading to the estimates in Table 3. While my substantive interest may be in how productivity affects tenure for men and women, gender differences in the probability of tenure for a given number of articles depends on whether a faculty member works in a distinguished department. To show this, Figure 6 plots gender differences in the probability of tenure for those not in distinguished departments while Figure 7 plots differences for those in distinguished departments. These figures show that gender differences in the probability of tenure are noticeably larger in distinguished departments for faculty with more than twenty-five articles. To combine the two figures and to introduce a technique that is essential for more complex models, I use a bold line to plot gender differences for those from distinguished departments and a thin line for those who are not. If the difference in probability is significant, the line is solid; if the difference is not significant, the line is dashed. Figure 8 shows how the prior graphs can be combined. With two predictors, not only isn't there a single test of the effect of articles, but now the tests for differences in probabilities depend on the level of articles and the prestige of the department. This complexity, however, may also provide interesting insights into the tenure process, suggesting that gender differences in the tenure process are greater in more prestigious departments.

[Table 4 and Figures 9, 10 & 11 about here]

As a more realistic example, I estimate the full model used in Allison (1999) with estimates shown in Table 4. Gender differences in the probability of tenure differ by the levels of five variables. Even with only five variables, which is a small number compared to the number of variables in many applications, it is impractical to examine all combinations of the variables. Instead, the examine predictions you must focus on those regions of the data space that are substantively most interesting. For example, in most universities tenure decisions are made in the seventh year in rank, so I compute predicted probabilities holding year at 7 and year-squared at 49.

As with Model 2, I am interested in the effects of articles and prestige. Since prestige is now measured as a continuous variable ranging from 1 to 5, I extend Figure 8 for Model 2 to show gender differences at five levels of prestige over a range of articles from 0 to 50. While I want to control for the selectivity of the baccalaureate, I am not particularly interested in its effect, so I hold this variables at the mean. Predictions are made by varying the number of articles and the level of prestige in these equations:

$$\begin{aligned}\pi^{M}\left(\text{articles, prestige}\right) &= \Lambda(\beta_0^M + \beta_1^M\left(7\right) + \beta_2^M\left(49\right) + \beta_3^M\overline{\mathit{selectivity}} \\ &\quad + \beta_4^M \mathit{articles} + \beta_5^M \mathit{prestige}) \\ \pi^{W}\left(\text{articles, prestige}\right) &= \Lambda(\beta_0^W + \beta_1^W\left(7\right) + \beta_2^W\left(49\right) + \beta_3^W\overline{\mathit{selectivity}} \\ &\quad + \beta_4^W \mathit{articles} + \beta_5^W \mathit{prestige})\end{aligned}$$

Varying articles for five levels of prestige leads to Figure 9.[7] Gender differences are smaller at lower levels of productivity, but the less prestigious the department the greater the difference. Since less prestigious departments are likely to require fewer publications for tenure than more prestigious departments, this is what would be expected. At higher levels of productivity, the differences are larger, with larger differences at more prestigious departments. This figure contains information on how both productivity and prestige affect tenure and our interpretation reflects the effects of both variables. To better understand how the effects differ by both prestige and productivity, I can plot the same predictions letting prestige vary continuously from 1 to 5 at fixed levels of productivity. This is shown Figure 10. In my experience, most people who first look at Figure 9 are surprised by Figure 10. When examining graphs of differences in predicted probabilities it is important to keep in mind that the curves are generated by the differences between two cumulative density functions that have different parameters. The complex nature of the nonlinearity is illustrated in Figure 11. In this three-dimensional figure, each dark or light band corresponds to .10 on the vertical axis. Broken lines indicate that the confidence interval at the prediction crosses 0. If you trace a curve as articles increase from 0 to 50 with prestige held at 5, you obtain the solid curve in Figure 9. Other curves from that figure are created by holding prestige constant at other levels. The curves in Figure 10 can be created in a similar fashion.

5 Conclusions

The comparison of groups in regression models for binary outcomes is complicated by the identification problem inherent in these models. Traditional tests of the equality of regression coefficients across groups confound the magnitude of the regression coefficients with residual variation. If the amount of variation differs between groups, the

[7] Color versions of these graphs are available at www.indiana.edu/~jslsoc/research_groupdif.htm.

test can lead to incorrect conclusions. The test proposed by Allison (1999) takes into account group differences in residual variation by incorporating the assumption that the effects of some variables are the same across groups. In practice, researchers might not have a basis for making this assumption, in which case the conclusions from this test could also be misleading. The approach advocated in this paper is to compare predicted probabilities across groups as levels of one variable change holding other variables constant. Since predicted probabilities are not affected by residual variation, testing the equality of predictions is not affected by the identification problem. Using predicted probabilities is *not*, however, without costs. Tests of the equality of regression coefficients across groups requires a single test and provides a simple answer to the question of whether a variable has the same effect for both groups. With predicted probabilities, you must examine group differences at many values to see how groups differ in the way that changes in predictors affect the outcome.

My proposal for comparing groups allows you to show if there are significant differences in the predicted outcomes and whether these differences vary by the level of the independent variables. This information does not correspond to a test of the equality of regression coefficients across groups. In models such as logit or probit, tests of group differences in probabilities do not allow us to determine whether these differences are due to differences between groups in their regression coefficients, to differences in residual variation, or to a combination of both. However, using probabilities allows us to show if there are significant differences in the predicted outcomes and whether these differences vary with the level of the independent variables. For many purposes, this information is critical for understanding the process being studied.

Using group differences in predictions requires a great deal of planning about which comparisons are most interesting and at which values you want to hold other variables in the model constant. As with logit and probit in general, the effect of a variable (in this case the effect of group membership) depends on the level of all variables in the model. Further, as you plan where you want to examine the predictions of your model, it is important to limit you explorations to those regions of the data space that are both substantively reasonable and where there are a sufficient observations to justify your conclusions. To illustrate this problem, consider Model 3 and the graphs showing gender differences in tenure by departmental prestige and the number of articles. While these differences are largest for those with high productivity in distinguished departments, the sample used was for scientists with degrees in the late 1950s and early 1960s. During this period there were few women in prestigious departments and hence the region of the data space where we find the largest gender differences has relatively few individuals. Still, the results are strong and consistent with historical findings about tenure at distinguished departments during this period (Rossiter 1982, 1995).

Finally, the approach advocated in this paper is not limited to binary logit and probit. The same methods can be used with ordinal and nominal models where similar issues of identification occur. Moreover, our approach is also valuable in models where it is possible to test regression coefficients across groups, such as a linear regression model. Examining group differences in the predicted outcome at specific values may provide insights that cannot be gained from a single test comparing coefficients.

Initial draft: February 2005
Revised draft: June 2009

6 References

Allison, Paul D. 1999. Comparing Logit and Probit Coefficients Across Groups. *Sociological Methods and Research* 28:186-208.

Cameron, A. C. and P. Trivedi. 2005. *Microeconometrics: Methods and Applications.* New York, Cambridge.

Chow, G.C. 1960. Tests of equality between sets of coefficients in two linear regressions. *Econometrica* 28:591-605.

Garfield, Eugene. 1984. This week's citation classic: "Chow, G.C. 1960. 'Tests of equality between sets of coefficients in two linear regressions.' *Econometrica* 28:591-605." *Current Contents* 49:16.

Liao, Tim Futing. 2002. *Statistical Group Comparison.* New York: Wiley.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, CA: Sage Press.

Long, J. Scott. 2009. GC: Programs for Group Comparisons of Regression Models for Binary Outcomes. Available running `findit scottlong` in Stata and following instructions to download the commands.

Long, J.S. and Freese, J. 2005. *Regression Models for Categorical and Limited Dependent Variables with Stata.* Second Edition. College Station, TX: Stata Press.

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.

Maddala, G.S. 1983. *Limited-dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

Rossiter, M.W. 1982. *Women Scientists in America: Struggles and Strategies to 1940.* Baltimore, Md., Johns Hopkins Press.

Rossiter, M.W. 1995. *Women Scientists in America: Before Affirmative Action 1940-1972.* Baltimore, Johns Hopkins Press.

StataCorp. 2007. *Stata Statistical Software. Release 10.* College Station, TX: Stata Corporation.

Williams, Richard. 2009. Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients Across Groups. *Sociological Methods & Research* 37(4): 531-559.

Xu, J. and J.S. Long. 2005. Confidence intervals for predicted outcomes in regression models for categorical outcomes. *The Stata Journal* 5: 537-559.

7 Appendix

The test in the paper is a special case of the more general test considered here. The logit or probit models are defined as

$$\Pr(y = 1 \mid \mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta})$$

where $F(.)$ is the cdf for the standard normal distribution for the probit model and the standardized logistic distribution for the logit model. Let $G(\boldsymbol{\beta})$ be the difference in the predicted probabilities at two sets of values of the independent variables contained in the vectors $\mathbf{x}_a$ and $\mathbf{x}_b$:

$$\begin{aligned} G(\boldsymbol{\beta}) &= \Pr(y = 1 \mid \mathbf{x}_a) - \Pr(y = 1 \mid \mathbf{x}_b) \\ &= F(\mathbf{x}_a'\boldsymbol{\beta}) - F(\mathbf{x}_b'\boldsymbol{\beta}) \ . \end{aligned}$$

This formulation is more general than the case considered in the text since it can be applied to any two sets of values of $\mathbf{x}$, not only comparisons between groups. For example, if the model is

$$\Pr(y = 1 \mid \mathbf{x}) = F(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$$

for a single group, $G(\boldsymbol{\beta})$ could be the difference in the predicted probability when $x_1 = 1$, $x_2 = 2$, and $x_3 = 2$ and when $x_1 = 2$, $x_2 = 3$, and $x_3 = 1$.

The paper estimates a model in which dummy variables and interactions are used to include both groups in the same equation. For example, consider a single independent variable x and two groups. Let $w = 1$ for women, else 0; let $m = 1$ for men, else 0; and let $wx = w \times x$ and $mx = m \times x$. Next, estimate the model

$$\Pr(y = 1 \mid \mathbf{x}) = \Lambda(\beta_w w + \beta_{wx} wx + \beta_m m + \beta_{mx} mx)$$

where no constant is included. For women, $w = 1$ and $m = 0$ which simplifies the equation to

$$\Pr\left(y = 1 \mid \mathbf{x}\right) = \Lambda\left(\beta_w + \beta_{wx}x\right)$$

and for men $w = 0$ and $m = 1$ so that

$$\Pr\left(y = 1 \mid \mathbf{x}\right) = \Lambda\left(\beta_m + \beta_{mx}x\right).$$

Group difference in the predicted probabilities at a given value of x equal

$$\begin{aligned} G\left(\widehat{\boldsymbol{\beta}}\right) &= \Pr\left(y = 1 \mid w = 1, m = 0, x\right) - \Pr\left(y = 1 \mid w = 0, m = 1, x\right) \\ &= \Lambda(\widehat{\beta}_w w + \widehat{\beta}_{wx} wx + \widehat{\beta}_m 0 + \widehat{\beta}_{mx} 0) - \Lambda\left(\widehat{\beta}_w 0 + \widehat{\beta}_{wx} 0 + \widehat{\beta}_m m + \widehat{\beta}_{mx} mx\right) \\ &= \Lambda(\widehat{\beta}_w + \widehat{\beta}_{wx} x) - \Lambda\left(\widehat{\beta}_m + \widehat{\beta}_{mx} x\right). \end{aligned}$$

Returning to the general formulation of the problem,

$$\frac{\partial G(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial F\left(\mathbf{x}_a' \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} - \frac{\partial F\left(\mathbf{x}_b' \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}}$$

where

$$\frac{\partial F(\mathbf{x}' \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial \Lambda(\mathbf{x}'\boldsymbol{\beta})}{\partial \beta_0} & \frac{\partial \Lambda(\mathbf{x}'\boldsymbol{\beta})}{\partial \beta_1} & \cdots & \frac{\partial \Lambda(\mathbf{x}'\boldsymbol{\beta})}{\partial \beta_K} \end{bmatrix}' = \lambda\left(\mathbf{x}' \boldsymbol{\beta}\right) \mathbf{x}.$$

Then,

$$\begin{aligned} Var\left[\Pr\left(y = 1 \mid \mathbf{x}_a\right) - \Pr\left(y = 1 \mid \mathbf{x}_b\right)\right] = & \left[\frac{\partial F(\boldsymbol{\beta} | \mathbf{x}_a)}{\partial \boldsymbol{\beta}'} Var(\widehat{\boldsymbol{\beta}}) \frac{\partial F(\boldsymbol{\beta} | \mathbf{x}_a)}{\partial \boldsymbol{\beta}}\right] \\ & + \left[\frac{\partial \Lambda(\boldsymbol{\beta} | \mathbf{x}_b)}{\partial \boldsymbol{\beta}'} Var(\widehat{\boldsymbol{\beta}}) \frac{\partial \Lambda(\boldsymbol{\beta} | \mathbf{x}_b)}{\partial \boldsymbol{\beta}}\right] \\ & -2\left[\frac{\partial F(\boldsymbol{\beta} | \mathbf{x}_a)}{\partial \boldsymbol{\beta}'} Var(\widehat{\boldsymbol{\beta}}) \frac{\partial F(\boldsymbol{\beta} | \mathbf{x}_b)}{\partial \boldsymbol{\beta}}\right]. \end{aligned}$$

The last term in the equation drops out when interactions are used to include both groups in the same equations and $\mathbf{x}_a$ and $\mathbf{x}_b$ are used to specify the values for each group. A $z-$statistic to test H_0: $\Pr\left(y = 1 \mid \mathbf{x}_a\right) = \Pr\left(y = 1 \mid \mathbf{x}_b\right)$ can be computed as

$$z = \frac{\Pr\left(y = 1 \mid \mathbf{x}_a\right) - \Pr\left(y = 1 \mid \mathbf{x}_b\right)}{\sqrt{Var\left[\Pr\left(y = 1 \mid \mathbf{x}_a\right) - \Pr\left(y = 1 \mid \mathbf{x}_b\right)\right]}}$$

which has an asymptotic normal distribution. Or, we can construct a confidence interval around $\Delta\left(\mathbf{x}_a, \mathbf{x}_b\right) = \Pr\left(y = 1 \mid \mathbf{x}_a\right) - \Pr\left(y = 1 \mid \mathbf{x}_b\right)$. For example, the 95% confidence interval for the difference in probabilities $\Delta\left(\mathbf{x}_a, \mathbf{x}_b\right)$ includes upper and lower bounds such that

$$\Pr\left(\Delta\left(\mathbf{x}_a, \mathbf{x}_b\right)_{\text{LB}} \leq \Delta\left(\mathbf{x}_a, \mathbf{x}_b\right) \leq \Delta\left(\mathbf{x}_a, \mathbf{x}_b\right)_{\text{UB}}\right) = .95.$$

Since the covariance term $\frac{\partial F(\boldsymbol{\beta}|\mathbf{x}_a)}{\partial \boldsymbol{\beta}'} Var(\widehat{\boldsymbol{\beta}}) \frac{\partial F(\boldsymbol{\beta}|\mathbf{x}_b)}{\partial \boldsymbol{\beta}}$ is zero when comparing groups, any software that computes confidence intervals and the standard error of the prediction can be used. For the more general case, the SPost commands (Long and Freese 2005; Xu and Long 2005) can be used for logit and probit as well as other models.

8 Tables and Figures

Table 1: Descriptive statistics for male and female biochemists using person-year observations.

		Women		**Men**		**Combined**	
	Variable	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Is tenured?	tenure	0.109	0.312	0.132	0.338	0.123	0.328
Year	year	3.974	2.380	3.784	2.252	3.856	2.303
Year-squared	yearsq	21.457	23.137	19.388	21.501	20.169	22.151
Bachelor's selectivity	select	5.001	1.475	4.992	1.365	4.995	1.407
Total articles	articles	7.415	7.430	6.829	5.990	7.050	6.576
Distinguished job?	distinguished	0.058	0.233	0.040	0.197	0.047	0.211
Prestige of job	prestige	2.658	0.765	2.640	0.784	2.647	0.777
N		1,056		1,741		2,797	

Table 2: Logit model 1 predicting tenure for male and female biochemists using only the number of articles.

	Women			**Men**		
Variable	β	exp (β)	z	β	exp (β)	z
constant	-2.501		-17.96	-2.721		-22.4
articles	0.047	1.048	4.49	0.102	1.102	9.76
log-likelihood	-353.758			-628.282		
N	1,056			1,741		

Table 3: Logit model 2 predicting tenure for male and female biochemists using articles and an indicator of having a high prestige job.

	Women			Men		
Variable	β	exp (β)	z	β	exp (β)	z
constant	-2.604		-17.32	-2.715		-22.27
articles	0.068	1.070	5.36	0.106	1.111	9.89
distinguished	-1.984	0.138	-2.69	-0.945	0.389	-2.06
log-likelihood	-348.344			-625.672		
N	1056			1,741		

Table 4: Logit models predicting tenure for male and female biochemists.

	Women			Men		
Variable	β	exp (β)	z	β	exp (β)	z
constant	-5.842		-6.75	-7.680		-11.27
year	1.408	4.087	5.47	1.909	6.745	8.92
yearsq	-0.096	0.909	-4.36	-0.143	0.867	-7.70
select	0.055	1.057	0.77	0.216	1.241	3.51
articles	0.034	1.035	2.69	0.074	1.076	6.37
prestige	-0.371	0.690	-2.38	-0.431	0.650	-3.96
log-likelihood	-306.191			-526.545		
N	1056			1741		

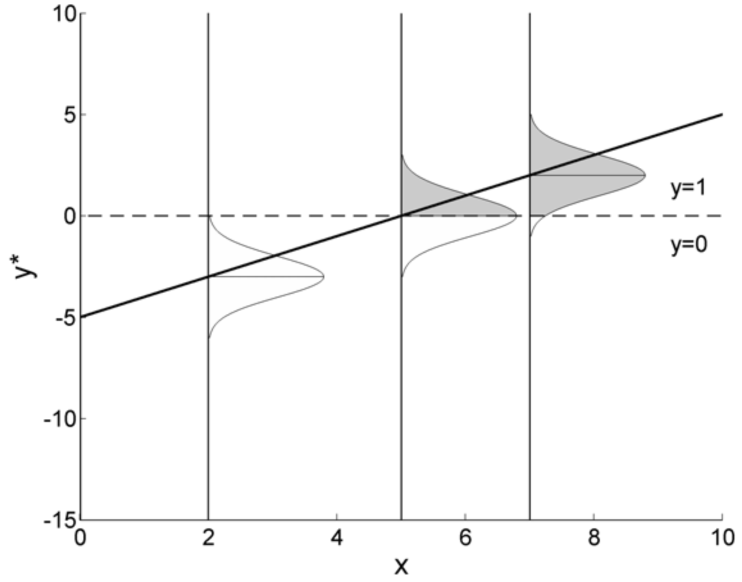


Figure 1: The link between $y^* = -6 + 1x + \varepsilon$ with $\sigma = 1$ and $\Pr(y = 1 \mid x)$ for parameter Set A.

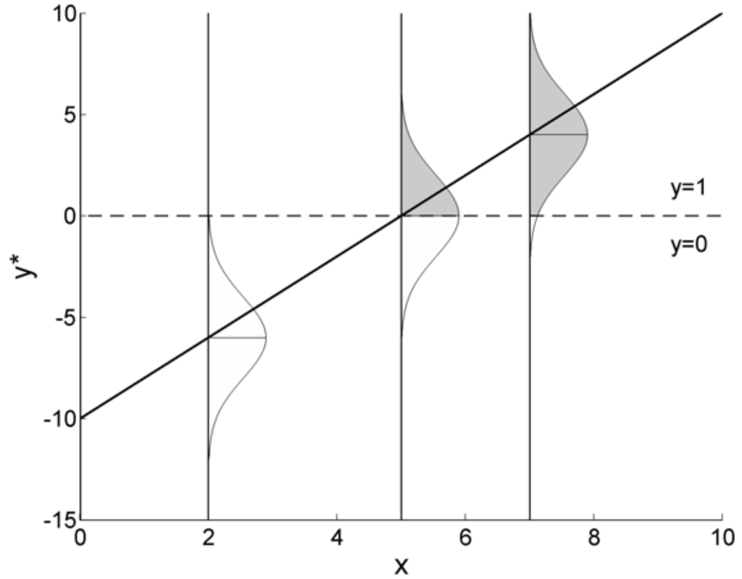


Figure 2: The link between $y^* = -12 + 2x + \varepsilon$ with $\sigma = 2$ and $\Pr(y = 1 \mid x)$ for parameter Set B.

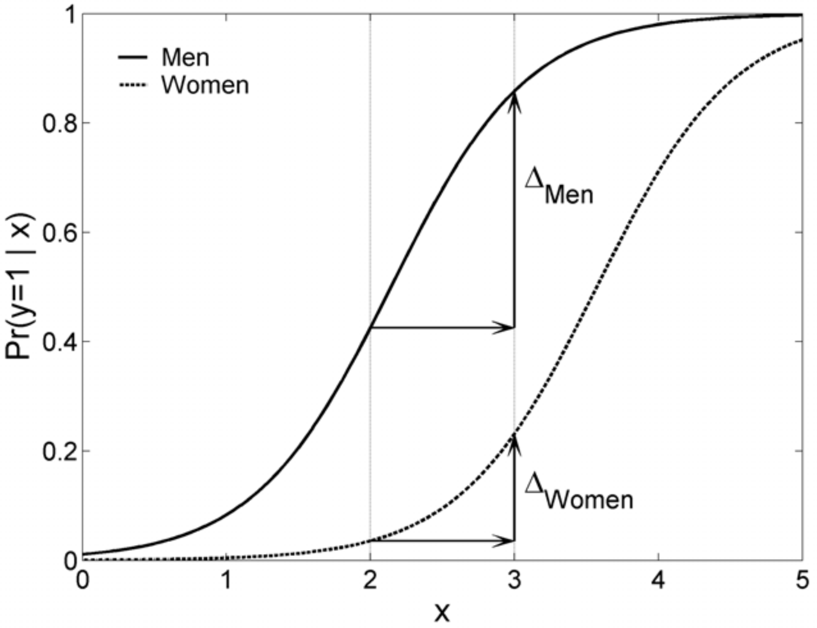


Figure 3: Changes in probabilities for men and women with the same coefficient for x but different intercepts.

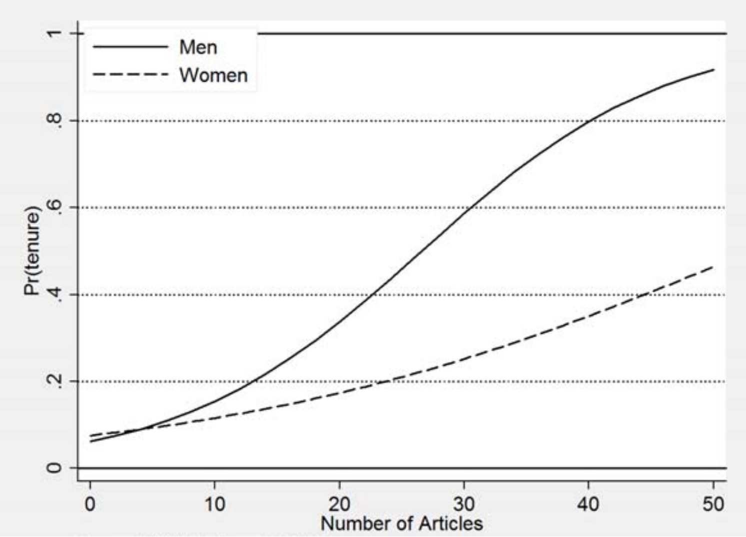


Figure 4: The probability of tenure for men and women by the number of articles using Model 1.

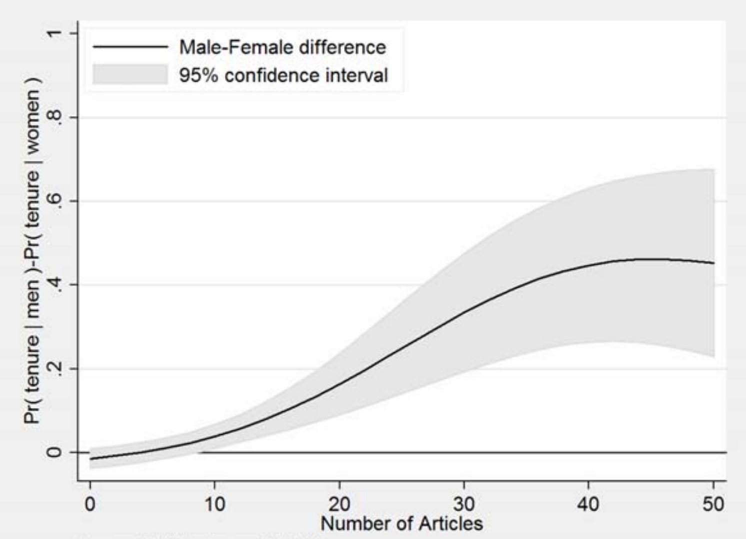


Figure 5: Gender differences in the probability of tenure by the number of articles using Model 1.

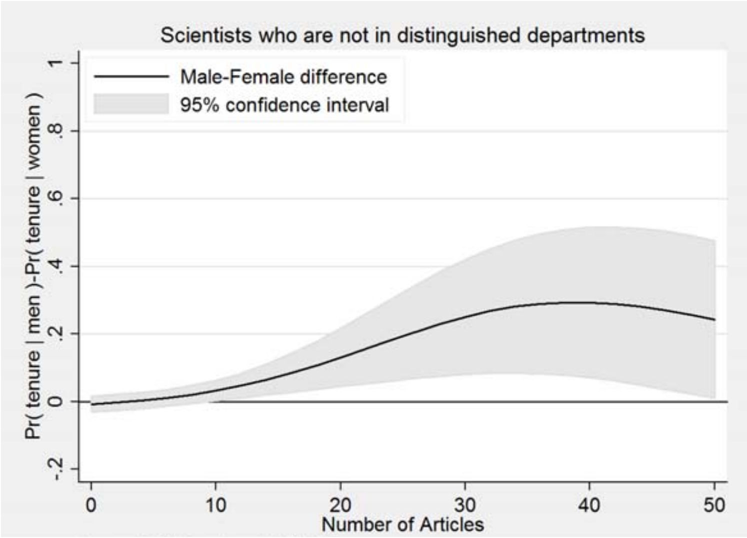


Figure 6: Gender differences in the probability of tenure by the number of articles for scientists who are not in distinguished departments using Model 2.

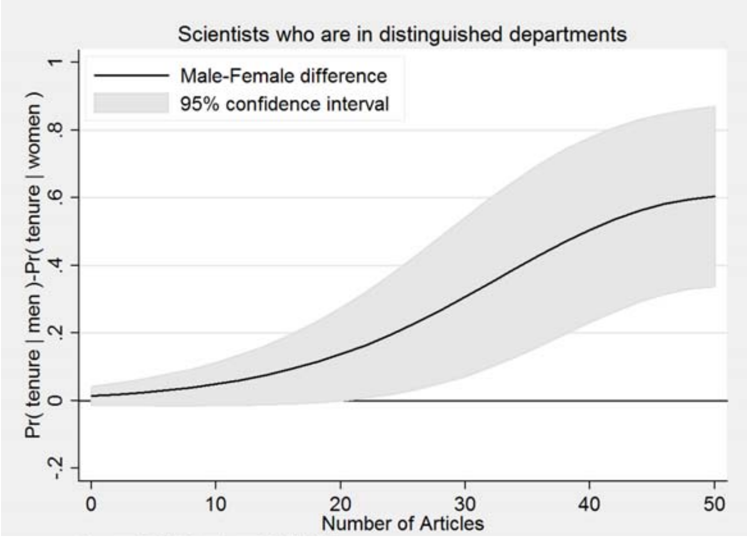
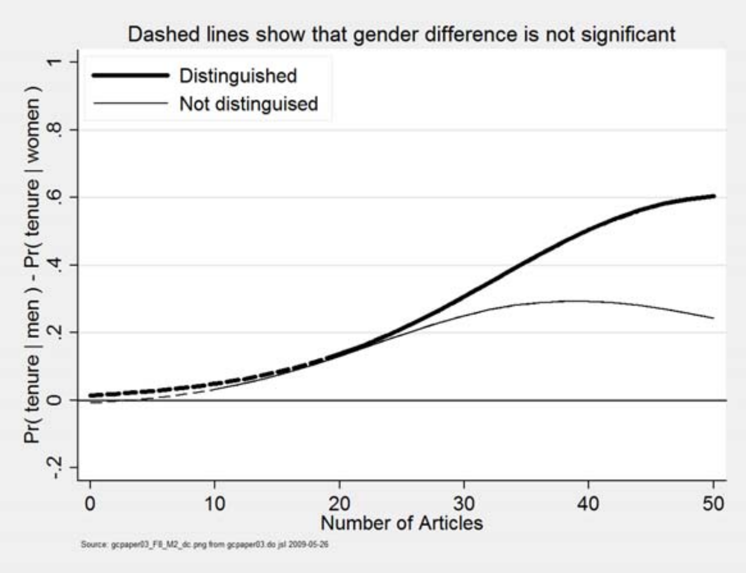


Figure 7: Gender differences in the probability of tenure by the number of articles for scientists who are in distinguished departments using Model 2.



Source: gcpaper03_F8_M2_dc.png from gcpaper03.do jsl 2009-05-26

Figure 8: Gender differences in the probability of tenure by the number of articles and by prestige of department using Model 2.

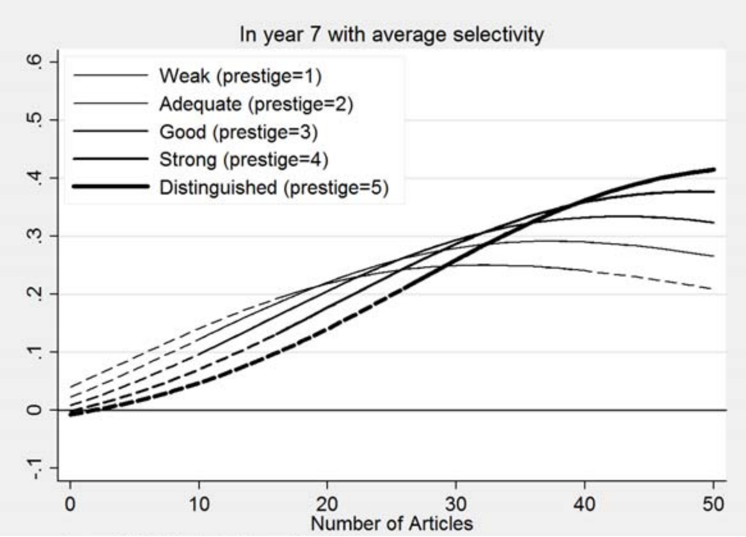


Figure 9: Gender differences in the probability of tenure by the number of articles at different levels of departmental prestige using Model 3.

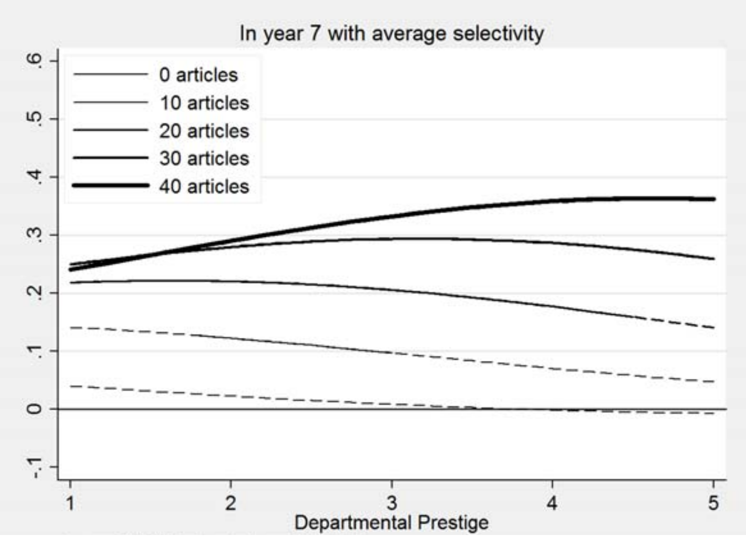


Figure 10: Gender differences in the probability of tenure by prestige at different levels of productivity using Model 3.

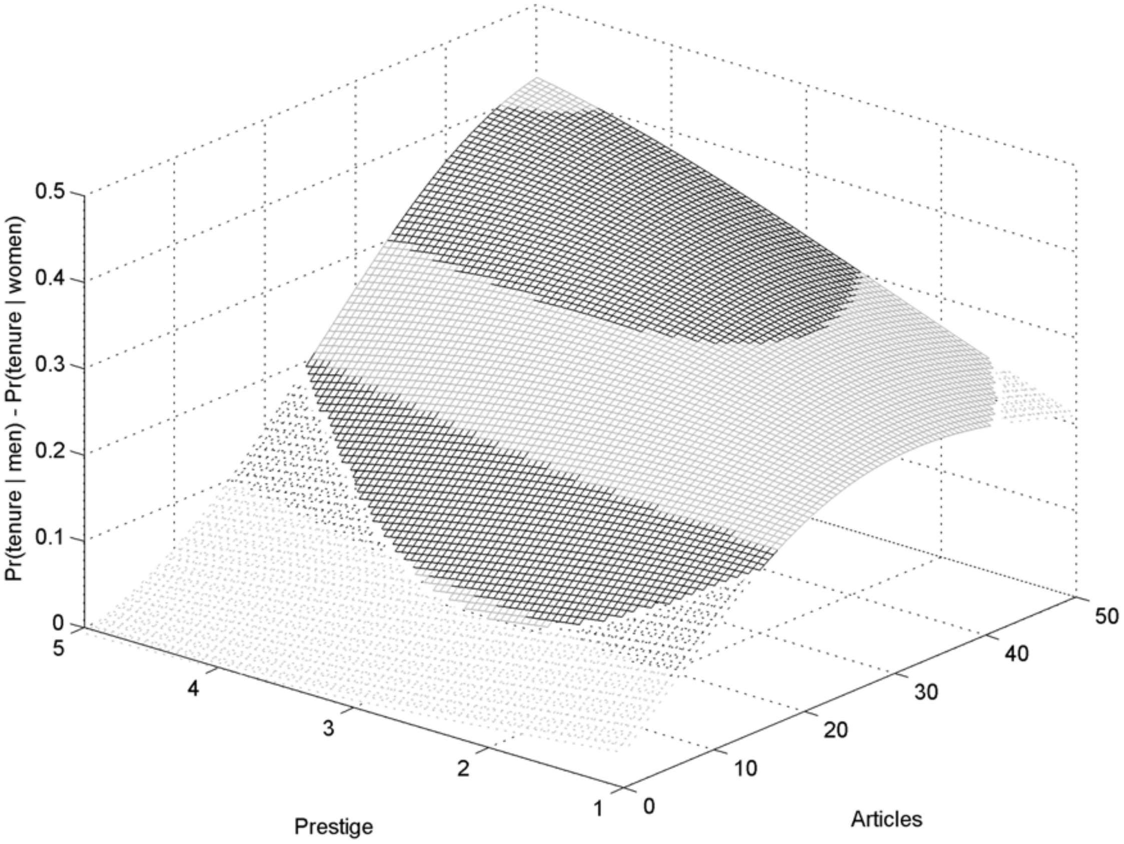


Figure 11: Gender differences in the probability of tenure by prestige and number of articles using Model 3.