CONFOUNDED COEFFICIENTS: EXTENDING RECENT ADVANCES IN THE ACCURATE COMPARISON OF LOGIT AND PROBIT COEFFICIENTS ACROSS GROUPS

GLENN HOETKER
College of Business
University of Illinois at Urbana-Champaign
350 Wohlers Hall, 1206 S. Sixth Street
Champaign, IL 61820
Phone: (217) 265-4081; Fax: (217) 244-7969; E-mail: ghoetker@uiuc.edu

Version: October 22, 2004

Acknowledgements: I gratefully acknowledge the helpful comments of Paul Allison, Jongwook Kim, Tim Liao, and Steve Michael. All errors remain my own.

Author's Biography

Glenn Hoetker is an assistant professor of strategy in the College of Business, University of Illinois at Urbana-Champaign. His research interests include social network analysis, the impact of national institutions on inter-firm relationships, and statistical methods. Recent publications include "Same rules, different games: variation in the outcomes of 'Japanese-style' supply relationships" (*Advances in International Management*, forthcoming) and "How much you know versus how well I know you: Selecting a supplier for a technically innovative component" (*Strategic Management Journal*, forthcoming).

CONFOUNDED COEFFICIENTS: EXTENDING RECENT ADVANCES IN THE ACCURATE COMPARISON OF LOGIT AND PROBIT COEFFICIENTS ACROSS GROUPS

ABSTRACT

The logit and probit models are critical parts of the sociologist's analytical arsenal. We often want to know if a covariate has the same effect for different groups, e.g., men and women. Unfortunately, many attempts to compare the effect of covariates across groups make the unwarranted assumption that each group has the same residual variation. If this assumption is false, comparisons of coefficients can reveal differences where none exist and conceal differences that do exist. Recent work has emphasized the theoretical potential for this problem and proposed a test of whether the effect of covariates differs across groups that is accurate, if limited, despite differences in residual variation. This paper extends these advances in three ways. First, it uses simulations to show that this theoretical problem is substantively significant under a wide range of common conditions, meaning that traditionally executed comparisons of logit coefficients should be viewed skeptically. Second, it uses simulations to assess the power of the test recently proposed to overcome the problem, finding that they are an improvement over naïve comparisons of coefficients, but have significant limitations. Third, it proposes and tests two alternative means of comparing coefficients across groups that avoid the assumption of equal residual variation entirely. The article closes with implications for the practice of research.

Keywords: Logit, probit, discrete choice

STATA CODE TO CARRY OUT THE TESTS DISCUSSED IN THIS ARTICLE IS AVAILABLE FROM THE AUTHOR.

INTRODUCTION

The logit and probit models have become critical parts of the sociologist's analytical arsenal. Researchers have used them to study topics ranging from support for capital punishment (Bailey 2002) to the likelihood that a particular venture capital firm invests in a given target company (Sorenson and Stuart 2001). Often it is of interest to know if a covariate has the same effect for different groups. For example, Long, Allison, and McGinnis (1992) studied whether the number of articles published affected the likelihood of promotion for male and female faculty equally and Pager (2003) asked if a criminal record had the same impact on the job prospects of blacks and whites.

Unfortunately, attempts to compare the effect of logit or probit coefficients across groups require an assumption that is often false. Logit and probit coefficients are scaled by the unknown variance of their residual variation. Naïvely comparing coefficients as one would in linear models assumes that residual variation is the same across groups, though in many cases it may not be. Differences in coefficients across groups may merely reflect the difference in residual variation across groups, rather than real differences in the impact of covariates across groups. As Allison (1999a:190) expresses it, "Differences in the estimated coefficients tell us nothing about the differences in the underlying impact of *x* on the two groups." Worse yet, comparisons may appear informative. They can reveal differences where none exist, conceal differences that do exist, and even indicate differences in the reverse direction of the actual situation.[1]

In a recent, influential paper, Allison (1999a) explicated the theoretical basis of this problem.[2] He then developed a set of related tests to determine if (a) the residual variation of two groups differs sufficiently to render traditional comparisons inappropriate, (b) if there is evidence that the effect of *at least one* covariate differs significantly across groups, and (c) if the effect of

[1] This same problem plagues cross-group comparison of coefficients in *any* generalized linear model, including logit, probit, and Poisson, as the variance of the dependent variable contains a scale factor which can vary across groups. Liao (2002:87-96) provides an extensive discussion of the topic, cast more broadly as dispersion heterogeneity in generalized linear models.

[2] An indication of the paper's influence is that 14 published articles have cited it in the four years since its publication.

a *specific* covariate differs across groups—assuming that the true effect of the other covariates is equal across groups. While Allison makes the theoretical basis for concern clear, it is less obvious under what conditions this theoretical concern will actually have substantive effect. Whatever the theoretical concerns, how likely are researchers actually to encounter situations in which the traditional means of comparisons would lead to the wrong conclusions?

The answer to this question is important, as neither of Allison's tests for the true effects of covariates differing across groups is completely satisfactory. A researcher must either be satisfied knowing that *at least one* covariate differs, but not which one(s), or base conclusions regarding a given covariate on the untestable assumption that the effect of the other covariates are identical across groups. Further, the tests require cannot be carried out with pre-packaged statistical routines. The programming required is straight-forward, but still adds a level of complexity to analysis.[3]

This paper contributes to the literature by extending Allison's advances in three ways. First, it uses simulations to explore the substantive significance of ignoring the assumption of equal residual variation, finding that even small differences in the groups' residual variation can make comparisons of coefficients extremely misleading. To my knowledge, this is the first time this has been demonstrated. Second, it uses simulations to ascertain the power of Allison's tests, finding that they are a significant improvement over naïve comparison of coefficients, although they have limitations and are not a panacea. Lastly, it proposes two additional approaches for comparing the underlying effect of covariates across groups that avoid the assumption of equal residual variation entirely.

[3] It is worth emphasizing that the algebra underlying the problem and potential solutions is straight-forward (see pages 192-4 of Allison (1999) for the latter). At issue is the importance of the problem in finite samples and the finite-sample performance of potential solutions. As Greene (2000:484) points outs, the small-sample properties are many test statistics are unknown, except in a few special cases. Prior work including Davidson & Mackinnon (1984), Orme (1995), Skeels and Vella (1999) and Yatchew and Griliches (Yatchew and Griliches 1985) has established that certain tests perform particularly poorly in finite samples and asymptotically equivalent variants of the same test can generate radically different small sample results. This is of particular relevance for logit and probit models, the finite-sample properties of which are largely unknown, unlike OLS models (Long 1997:53).

The article proceeds by briefly reviewing the logit and probit models, particularly the potential problem caused by inter-group differences in residual variation. I then present the simulations used to explore the problem and Allison's approaches to addressing it. Next, I propose two alternatives that avoid the assumptions of equal residual variation entirely. The paper concludes with a discussion of implications for the practice of research using logit and probit.

IMPLICATIONS OF ASSUMPING EQUAL RESIDUAL VARATION ACROSS GROUPS

Statistical implication of the assumption

Since standard econometric texts (e.g., Greene 2000) and more specialized works (Maddala 1983; Train 1986; Long 1997; Allison 1999b) cover the logit and probit models, this section will review the logit model only briefly, focusing on the elements relevant to this paper. As much of the paper builds on Allison (1999a), I closely follow his exposition for the convenience of the reader.

Suppose we are modeling which of two alternatives occurs, e.g., whether a faculty member achieves tenure. Without loss of generality, we assign a value of 0 to the dependent variable y_i for cases in which one alternative occurs and set y_i equal to 1 for cases in which the other alternative occurs. We assume that y equals 1 only if an unobserved, continuous variable y^* is greater than an unobserved threshold, τ. That is,

$$y_i = \begin{cases} 1 \text{ if } y_i^* > \tau \\ 0 \text{ if } y_i^* \leq \tau \end{cases} \qquad (1)$$

Further, we assume that y^* is linearly related to the observed independent variables:

$$y_i^* = \boldsymbol{\alpha}\mathbf{x}_i + \sigma\varepsilon_i \qquad (2)$$

where $\mathbf{x_i}$ is a vector of observed covariates and ε_i is a random disturbance independent of the observed covariates. As in the linear model, the disturbance reflects the impact of differences across cases in variables the researcher does not observe—residual variation. σ is a scale parameter that allows the amount of residual variation to be greater or smaller.

Since y^* is a latent variable, we cannot estimate its variance. In the logit model, we assume ε has a logistic distribution (also called a Gumbel or type 1 extreme value distribution) and variance $\pi^2/3$. In the probit case, we assume ε is normally distributed with a variance of 1. These arbitrary values are assigned as identifying assumptions and cannot be confirmed by the data.

Letting p_i represent $\Pr(y_i=1|x_i)$, these assumptions lead to the familiar logit model[4]

$$\ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x_i}\boldsymbol{\beta} \qquad (3)$$

The β coefficients we obtain from estimating equation (3) are related to the α coefficients in equation (2) as follows.

$$\beta = \frac{\alpha}{\sigma}. \qquad (4)$$

This relationship is the heart of the problem in comparing coefficients across groups. Since the value of σ is unidentifiable, we cannot recover α. The unobservable value of σ determines the scale of β. Accordingly, if σ varies between groups, the logit coefficient β will also vary, *even if the underlying effect of the covariate on y^*, α, is the same between groups.*[5]

Substantive implications of the assumption

In many contexts, it is reasonable to assume that residual variation differs across groups. For example, institutional pressures may lead Japanese firms to be more similar in their strategy than U.S. firms (Lincoln 2001). In labor mobility studies, there is evidence that women have more heterogeneous career paths than men (Long and Fox 1995).

There is a heavy burden of proof on any author claiming that residual variation is the same across groups, because the failure of this assumption can have serious consequences. To demonstrate these consequences, consider the following hypothetical model, which uses a simulated dataset. We wish to model the dichotomous variable *y* as a result of two independent variables, x_1 and x_2. We are particularly interested in knowing whether the effect of x_2 on the

[4] Since the same development applies to both the logit and probit models, I subsequently limit my discussion to the logit for simplicity.

[5] Within a group, equation 4 poses no problem. Whatever the value of σ, β will only be zero when α is zero. Tests of the significance of β will therefore provide accurate inferences of the relationship between associated covariate and y*.

likelihood of y=1 differs across two groups, which I unimaginatively label "Group 0" and "Group 1". I generated the data according to equation (5).[6]

$$\begin{aligned} y_i^* &= x_{1i} + 2x_{2i} + \sigma_g \varepsilon_i \\ x_1, x_2 &\sim \mathrm{N}(0, 4) \\ \sigma_1 &= 2 \\ \sigma_2 &= 4 \end{aligned} \tag{5}$$

Note that the actual impacts of x_1 and x_2 are the same for both groups. Only the residual variation differs across groups. For simplicity, I assume the two groups are the same size and that group membership is exogenous, e.g., gender or firm's home country.[7]

There are two common approaches to comparing coefficients across groups: comparing the coefficients that result from estimating separate models for each group or estimating a single model that interacts a variable for group membership with variables of interest. Table 1 reports the results of the first approach. I estimated the following equation twice, once for each group.

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{6}$$

Naively comparing the results, we see an apparent difference in the impact of x_2. It appears to have almost twice as strong an effect (1.835/.922) for Group 0 as for Group 1. In a linear model, we would use the Wald chi-squared statistic to determine if the difference in the estimated coefficients is statistically significant. Assuming the coefficients for each group have independent sampling distributions, the statistic is

$$\frac{\left(\beta_2^{G1} - \beta_2^{G0}\right)^2}{\left[std.err.\left(\beta_2^{G1}\right)\right]^2 + \left[std.err.\left(\beta_2^{G2}\right)\right]^2}, \tag{7}$$

[6] All data generation and estimation was performed using Stata version 8/SE. To ensure replicability, random number generation was initiated with a seed of 125. Logistic error terms were generated from $\ln\left(\ln\left(1/1-U(0,1)\right)\right)$, where U(0,1) indicates a random number from the uniform distribution. (See http://www.resacorp.com/gumbel_random.htm). y_i was set to 1 if $y_i^* > -.577$. (Textbook presentations normally express the threshold as 0, which is of course completely arbitrary. The mean of the extreme value distribution is actually *not* zero (Train 2004:39), which is immaterial to the estimation. Setting the threshold at the mean value as I have done here simply means that approximately 50% of the observations have y_i=1.)

[7] If group membership is endogenously determined, e.g., firms' choice of entry mode into a foreign market, selection bias must be addressed. See Shaver (1998) for details.

which has one degree of freedom. Applying it to the coefficients for x_2 yields a statistic of 31.12, which is highly significant ($p<.001$). We would thus conclude that the effect of x_2 differs across groups. This conclusion is, of course, false, since by construction the only difference between the two groups is the scale of their residual variation. Unfortunately, in a real world application, we would not know the residual variation in each group and could not tell if differences in coefficients indicate differences in actual effect.[8]

Table 2 extends this simulation to show how severe the problem can become with even small differences in the residual variation. I again generated data using Equation (5), but let σ_1, the scale parameter for Group 1's residual variation, range from 1 to 2 in increments of 0.2. I generated 1000 datasets of 1000 observations each for the five different values. The table shows the results of estimating Equation (6) on each dataset. The first column shows that the estimated value of β_2 decreases quickly as the residual variation increases, even though the value of the α_2 coefficient is exactly the same in each dataset. The next column uses Equation (7) to test for the equality of β_2 between Group 1 and Group 0 (for which the scale of the residual variation was held at 1.0). When σ_1 was 1.2, the coefficients were incorrectly found to differ at the five percent level of significance in 223 out of 1000 cases. Larger differences in the residual variation led to more false results, as one would expect. For example, when σ_1 was 1.6, the coefficients were found to differ in 923 out of 1000 cases. This clearly illustrates the hazards of comparing coefficients across groups if their residual variation might differ. The theoretical concerns are substantively relevant given even relatively small differences in residual variation across groups.

[8] Odds-ratios are also often used in interpreting logit coefficients. In the context of cross-group comparisons, they present several particular challenges. First, since they are merely a transformation of the estimated β_k coefficient, they remain susceptible to the problems discussed in this article. Long (1997:82) points out that the change in probability implied by a given change in odds depends critically on the current odds; these may well vary across groups, creating an additional pitfall for cross-group comparisons. Comparisons of the change in probability at specific values of the covariates are valid, although these chosen values must be valid for both groups if the comparison is to be meaningful.

Changes in predicted probabilities at theoretically relevant values of covariates can also be calculated. Again, the values chosen must be valid for both groups if the comparison is to be meaningful. For either approach, the issue of determining the statistical significance of any differences found remains. Therefore, this paper will focus on comparison of coefficients.

This finding means that traditionally executed comparisons of coefficients should be viewed with skepticism.

The second approach to comparing the effect of a covariate across groups is to estimate a single regression for all observations (Aiken and West 1991; Pindyck and Rubinfeld 1991). Although Allison (1999a) does not discuss this approach, it is common enough to merit examination. A dummy variable is set to one for one type of observation, e.g., Group 1, and interacted with the relevant covariates. Considering a linear model with only one variable, *x*, and letting the dummy variable, G, be set to 1 for Group 1 firms, the equation would be of the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (G_i x_i) + \varepsilon_i \qquad (8)$$

An estimate of β_2 significantly different from 0 indicates that the impact of *x* varies between Group 0 and Group 1. The sign of β_2 indicates whether the impact of *x* is diminished or increased for Group 1.

As indicated by the presence of a single error term, ε_i, this approach assumes that the residual variation for Group 0 and Group 1 is the same (Pindyck and Rubinfeld 1991:107; Darnell 1994:111). We can test this assumption in the linear model (Quandt 1960), but as discussed above, we cannot test it in a logit model, because the standard deviation of the residual variation cannot be identified and has been arbitrarily set to $\pi^2/3$ (Maddala 1983:23; Long 1997:47). Therefore, the single equation approach is inappropriate unless there are strong theoretical reasons to believe that the residual variation is the same in both groups.[9]

A variation of the above simulation demonstrates how misleading this sort of comparison can be. I generated 1000 simulated datasets of 1000 observations each, 500 each from Group 0 and Group 1, according to the following equation.

[9] Ai and Norton (2003) discuss other important issues in the use and interpretation of interaction terms in non-linear models, even if the assumption of equal residual variation holds.

$$y_i^* = x_{1i} + 2x_{2i} + .5(G_i x_{2i}) + \sigma_g \varepsilon_i$$
$$G_i = \begin{cases} 0 \text{ for group } 0 \\ 1 \text{ for group } 1 \end{cases} \qquad (9)$$
$$x1,\ x2 \sim N(0, 4)$$
$$\sigma_0=1;\ \sigma_1=3$$

I then estimated the equation

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (G_i x_2) \qquad (10)$$

on each dataset. A significant coefficient for β_3 would indicate that x_2's impact on the dependent variable differs across groups. Even though x_2's effect is 25% greater for Group 1 by design, estimating Equation (10) fails to find this difference. Figure 1 shows the point estimate and 95% confidence intervals for β_3 in each of the first 50 of 1000 simulated datasets. The 95% confidence interval included the true value of the coefficient, 0.5, in only two datasets: 4 and 23. However, the estimated coefficient was not significant in either dataset. Disturbingly, the only statistically significant estimates, datasets 5, 8, and 26, were each *negative*, indicating a diminished impact of x_2, rather than the true increased impact. This pattern is replicated in the full set of simulated datasets. β_3 was properly estimated as positive and significant in only 4 of 1000 datasets, while it was incorrectly identified as significant and negative in 94 datasets.

This unfortunate outcome is not surprising. The likely outcome of using an interaction term in a single equation despite differences in variance between the groups is that the slope coefficients will be found *not* to differ, even if they actually do (Gujarati 1988: 527). However, datasets 5, 8, and 26 show that it is also possible to find an effect contrary to reality. Clearly, the single equation with interaction term strategy is no better than estimating each group separately.

In fact, it is even less informative. In the single equation model, it is impossible to determine if x_2 has a significant effect within group 1. Because its effect is $\beta_2 + \beta_3$, calculations of its effect are tainted by the errors in estimating β_3. In contrast, the separate equations reported in Table 1 show that x_2 had a statistically significant effect for both Group 0 and Group 1.

To summarize, we cannot compare coefficients across groups in a logit or probit model as we would in a linear regression. Doing so, either by regressing each group separately and comparing coefficients or by using an interaction term in a single equation, can lead to erroneous conclusions. Insignificant differences can appear significant and significant differences can appear insignificant. Clearly, we need a way to identify a difference in residual variation across groups and to carry out accurate comparisons even in its presence.

IDENTIFYING AND ADJUSTING FOR UNEQUAL RESIDUAL VARIATION

Allison's method

Allison (1999a) developed a set of related tests to determine if (a) the residual variation of two groups differs significantly, (b) if there is evidence that the true effect of *at least one* covariate differs significantly across groups, and (c) if the true effect of a *specific* covariate differs across groups. I will only briefly sketch the underpinnings to this method, focusing on demonstrating its power and relative ease of application. This section introduces Allison's method and uses simulations to test its power.[10]

At the heart of the approach is rewriting the underlying model as a single equation that allows the residual variation to vary across groups. Under the assumption that all coefficients are equal across the groups, we can write the model as

$$
\begin{aligned}
y_i^* &= \beta_0 + \beta_1 G_i + \sum_{j>1} \beta_j x_{ij} + \sigma_i e_i \\
G_i &= \begin{cases} 0 \text{ for group } 0 \\ 1 \text{ for group } 1 \end{cases} \\
\sigma_i &= \frac{1}{1+\delta G_i}, \quad \delta > -1
\end{aligned}
\tag{11}
$$

[10] Allison briefly discusses extending his method to more than two groups, allowing an investigator to determine if the effect of a given variable is constant across all groups. I will limit my discussion to the two group case. Liao (2002:89-90) presents a generalized F-test that allows the research to test for the equality of residual variation across multiple groups. Liao notes that tests of equal residual variation for more than two groups may not give the same results of pairwise tests, which rely on the particular pairs of groups chosen for comparison.

Arbitrarily setting σ equal to 1 for group 0, δ>0 implies that the residual variation is smaller for Group 1 than Group 0. If δ<0, the residual variation is larger for Group 1 than Group 0. The standard deviation of the residual variations differs by 100δ percent.

Combining this with Equation (3) leads to

$$\ln\left(\frac{p_i}{1-p_i}\right) = \left(\beta_0 + \beta_1 G_i + \sum_{j>1}\beta_j x_{ij}\right)(1+\delta G_i), \qquad (12)$$

which can be estimated using code supplied by Allison.[11]

The first test proceeds under the null hypothesis that the values of the underlying coefficients are the same across groups, but that the residual variation differs. That is, it tests that the α terms are the same, but that σ varies across groups. The test proceeds by estimating equation (12) and examining $\hat{\delta}$. We can determine if $\hat{\delta}$ is significantly different from zero by a Wald chi-square test (the squared ratio of the estimate to its standard error). Alternatively, we can construct a log-likelihood ratio test by taking twice the positive difference between the log-likelihood for this model and the log-likelihood for an ordinary logit equation (equivalent to assuming δ=1). If $\hat{\delta}$ is not significant, the test provides no evidence that the residual variation differs between groups. In this case, Allison suggests continuing with conventional methods for comparing coefficients.

If, on the other hand, $\hat{\delta}$ is significantly different from zero, it is evidence that the residual variation differs across groups. The residual variation differs by $100\hat{\delta}$ percent between groups, with a positive value indicating that Group 1's residual variation is greater than Group 0's.

If we find unequal residual variation, the next step is to test the null hypotheses that the α coefficients are the same across groups versus the alternative hypotheses that at least one of them varies. Since the model estimated immediate above constrains the α terms to be equal across groups, we need to compare it to an unconstrained model that allows the α coefficients to vary across groups. We do so with a likelihood ratio test. We can obtain the log-likelihood for the unconstrained model by adding together the log-likelihoods obtained by estimating a separate logit model for each group.

[11] In Allison's original code, replace "$ml_y1" with "$ML_y1", noting the capitalization.

I return to the earlier simulation to explore the power of this method. To improve on conventional practice, it should meet three criteria. First, it should reveal when residual variation differs significantly between groups. Second, it should detect when apparent differences in coefficients across groups are merely the impact of differences in residual variation. Third, it should allow us to detect true differences of coefficients across groups.

I first test the method's performance when the underlying coefficients, the α's in Equation (3), are the same and only the residual variation varies across groups. I generated datasets of 1000 observations (500 of Group 0, 500 of Group 1) according the following equation.[12]

$$\begin{gathered} y_i^* = x_{1i} + 2x_{2i} + \sigma_g \varepsilon_i \\ x_1, x_2 \sim \mathrm{N}(0, 4) \\ \sigma_0=1, \sigma_1=1, 1.2, 1.4, 1.6, 1.8 \end{gathered} \qquad (13)$$

I varied the scale of Group 1's residual variation from 1 to 1.8. At each level, I generated 1000 datasets.

I then tested for differences in residual variation by estimating equation (12) on each dataset. The first two columns of Table 3 report the results. The method does moderately well in meeting the first criteria, detecting differences in residual variation between groups. When the scale of Group 1's residual variation is 1.8 times that of Group 0, $\hat{\delta}$ is significantly different from zero in 933 out of 1000 cases (880 out of 1000 using a likelihood ratio test). That is, the method accurately indicated a difference in residual variation in the overwhelming majority of cases. When the difference is more moderate, σ_1=1.4, both the Wald and log-likelihood ratio tests indicate a significant difference in residual variation in only about half of the simulated data sets.[13] Importantly, the test yields relatively few false positives. When the residual variation was

[12] The number of observations should not affect differences in residual variation, as the issue is one of identification, which additional data cannot address. However, sample size does affect our ability to assess accurately differences in unobserved heterogeneity and underlying coefficients, since the precision with which test statistics are estimated is influenced by sample size. The simulations in this paper hold the number of observations constant to focus on changes in other aspects of the estimation. However, unreported simulations indicate the smaller samples are considerably less powerful. For example, Allison's method is up to seven times *less* powerful at detecting true differences in coefficients (shown in Table 4) with a sample of 100 observations, rather than 1000.

[13] Recall that a 40% difference in the scale of residual variation caused conventional tests to falsely indicate a difference in the true effect of a covariate in 681 of 1000 cases. Allison commented (personal communication, December 17, 2002) that the simulation results "call into question my recommendation" to proceed with standard means of comparing coefficients if the test fails to reject the hypothesis that the groups have equal residual variation.

equal across groups, the method indicated so in 930 out of 1000 cases (Wald test). That is, it falsely indicated that residual variation differed across groups in only 70 of the 1000 cases.

The third column shows the results of testing the null hypothesis that the true coefficients are all equal across groups versus the alternative that at least one differs. Allison's test correctly indicated that there is no actual difference in the coefficients in approximately 950 out of 1000 cases at each level of Group 1's residual variation. That is, it incorrectly rejected the null hypothesis of no difference in only approximately 50 of 1000 cases. To emphasize the improvement over conventional tests, compare the results to Table 2. There, the conventional test incorrectly indicated that x_2's effect differed across groups in 223 out of 1000 cases when Group 1's residual variation was 1.2 times greater than Group 0's, and in 681 out of 1000 cases when Group 1's residual variation was 1.4 times greater. Clearly, the test meets the second criteria for improving on current practice.

I next test the method's ability to detect *true* differences in the value of a coefficient across groups, the third criteria. I again generated datasets of 500 Group 0 and 500 Group 1 observations, this time using Equation (14). It fixes the scale of Group 1's residual variation slightly higher than that of Group 0 and varies γ, the additional impact of x_2 for Group 1, from 0 to 1 in 0.2 increments. Since the coefficient for x_2 is 2 for Group 0, this range represents an up to fifty-percent greater effect for Group 1. For each value of γ, I generated 1000 datasets.

$$
\begin{aligned}
y_i^* &= x_{1i} + (2 + \gamma G_i)x_{2i} + \sigma_g \varepsilon_i \\
G_i &= \begin{cases} 0 \text{ for group } 0 \\ 1 \text{ for group } 1 \end{cases} \\
&\text{x1, x2} \sim N(0, 4) \\
&\sigma_0 = 1;\ \sigma_1 = 1.4 \\
&\gamma = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1.0
\end{aligned}
\qquad (14)
$$

The first two columns of Table 4 report on the method's ability to detect differences in the residual variation across groups. When x_2 has the same effect in both groups, the method

---

Further research on this point is needed. A conservative course of action would be to test the null hypothesis that none of the true coefficients vary across groups, even if the method does not indicate a difference in the residual variation of the groups.

accurately reports that the residual variation differs in just over half the sample datasets. Unfortunately, the method becomes *less* able to detect the difference in residual variation across groups as x_2's additional impact on Group 1 increases. When x_2 has a 20% greater impact on Group 1 (γ=.4), the method identifies the difference in residual variation in only 249 of 1000 cases. Since we know from Table 2 that this degree of difference in residual variation can lead to false conclusions about differences in coefficients, this result raises concerns.

The third column of the table reports on the method's ability to detect real difference in coefficients across groups. When x_2 has a 10% greater impact on Group 1 (γ=.2), the method reports the difference in only 46 of 1000 cases. However, when the difference is 20%, (γ= .4), it detects the difference in 619 of 1000 cases. It continues to improve as the difference increases, as one would expect.

It is also theoretically possible to test the null hypothesis that all of the underlying coefficients are the same against the alternative hypothesis that a *specific* coefficient, e.g., x_2, differs across groups. However, the test assumes that all of the coefficients not being tested are the same across groups. We cannot test this assumption without engaging in circular logic. To test whether the coefficients for x_2 differ, we must assume that the coefficients for x_1 are the same. However, we cannot test that assumption without assuming that the coefficients for x_2 are the same. Given this limitation, the scope for applying this test is limited and I will not test its power.

In general, simulation results are somewhat reassuring about our ability to detect and resolve the confounding effect of different residual variation across groups. However, even when the difference in the scale of the residual variation was forty percent, Allison's method failed to indicate this difference in almost half of the cases. Still, by identifying the difference in even half the cases, it greatly improves on naïve comparison of coefficients and should be routinely

applied. However, it is not a panacea. Therefore, I will present two alternative approaches that avoid assumption of equal residual variation entirely.[14]

Alternative 1: Differences in the relative effect of covariates

Given the limitations of the method above, particularly its inability to show whether a *specific* coefficient differs across groups, we may want to consider an approach that renders any difference in residual variation irrelevant. Suppose we could frame our interest not as whether the absolute effect of x_2 differed across groups, but rather as whether the impact of x_2 relative to x_1 differs across groups. To answer this question, we compare the ratio β_2/β_1 (Train 1998:237). If this ratio were 2 for Group 0 and 3 for Group 1 it would mean that a "unit" of x_2 has twice the effect of a unit of x_1 for Group 0 and thrice the effect of a unit of x_1 for Group 1. Relative to x_1, x_2 has a stronger effect on Group 1.

When this sort of comparison is sensible and theoretically interesting, the nature of ratios provides us a powerful benefit. Since β is the underlying coefficient, α, divided by the scale of the residual variation, σ, we find that

$$\frac{\beta_2}{\beta_1} = \frac{\alpha_2 / \sigma}{\alpha_1 / \sigma} = \frac{\alpha_2}{\sigma}\left(\frac{\sigma}{\alpha_1}\right) = \frac{\alpha_2}{\alpha_1}. \qquad (15)$$

By taking a ratio, we have removed the impact of residual variation and are left with a ratio of the underlying effects of x_2 and x_1. We can compare this ratio across groups, since it is no longer confounded by differences in residual variation.

The statistical significance of the difference in the ratios across groups can be computed with a Wald chi-squared test (Greene 2000).[15] Unfortunately, even large differences in ratios may not be statistically significant, especially if one or more terms are estimated with poor precision. To demonstrate this, I generated data according to Equation 14 above and then compared the ratio of β_2 to β_1 resulting from estimating

[14] It is also possible to deal with unequal residual variation more directly via double generalized linear models (DGLMs) and mean and dispersion additive models (MADAMs). See Liao (2002:91-92) for further information and references to the necessary code in S and Glim respectively.

[15] Code necessary to carry this test, and the others discussed in this paper, is available from the author.

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad (16)$$

separately for each group.

Table 5 shows the results given sample sizes of 100 and 1000. When the true coefficient of x_2 is 2.4 for Group 1 versus 2.0 for Group 0, the difference in the ratio β_2/β_1 across groups is statistically significant in 614 of 1000 simulated datasets of 1000 observations. With only 100 observations, however, we detect the difference in only 411 of 1000 datasets. The difference becomes more extreme when the true value of x_2 rises to 2.8 for Group 1. With 1000 observations, we detect the difference in β_2/β_1 across groups in almost every case, 983 out of 1000. With only 100 observations, however, we detect the difference in only 590 out of 1000 cases. Clearly, we need more precise estimates and thus more observations to apply this technique than we need to compare individual coefficients.

To provide concrete examples of how one might apply this technique, I draw upon three articles from recent issues of the *American Journal of Sociology*. While the articles did not necessarily apply the logit model, they illustrate the potential application of this technique in settings of interest to sociological researchers.

Simon and Nath (2004) model the expression of negative or positive emotion as a function of covariates including having children and household income. Differences in social expectations might suggest a difference in the relative importance of each factor for men and women, which could be tested by comparing the ratio of the corresponding coefficients.

Zeng and Xie (2004) model the earnings of U.S. and foreign-educated Asian-Americans as a function of covariates including job experience and years of education on earnings. One might hypothesize that, relative to job experience, employers value U.S. education more than foreign education, since they are less familiar with the reputations of foreign educational institutions, while job experience is more easily understood. This hypothesis could be easily tested by comparing the ratio of the coefficients for education and job experience for U.S-educated and foreign educated Asian-Americans.

Lastly, Gangl (2004) studied the probability that unemployment would lead to a significant loss of earnings in subsequent jobs in the U.S. and former West Germany. Cross-national differences in the relative importance of years of education and the availability of unemployment insurance has immediate policy implications in terms of the most effective response to large-scale unemployment—should displaced textile workers be retrained or should they be given extra support to extend the period of time they can search for the optimal replacement job? Because of differences in labor market rigidity, the optimal policy response could well vary across nations.

These examples demonstrate the range of issues that can be investigated by comparing ratios of coefficients. Of course, researchers are responsible for making sure that the scales involved make sense. Comparisons across studies are especially likely to be vulnerable to differences in measurement scales.

Alternative 2: Abandon direct comparisons

Given these challenges, a researcher may wish to simply abandon direct comparisons of coefficients across groups. Even in this case, we can often make some analytical progress.

If we model the two groups separately, the coefficients and standard errors are consistent *within each group*. The pattern of coefficient significance between the two models may provide some information. If β_1 were positive and highly significant for Group 0 and far from significant for Group 1, it would be informative to report that x_1 was significant for Group 0, but not for Group 1. Obviously, this statement is more informative if the samples are of roughly the same size, the model appears well specified and the *p*-values do not straddle a particular significance level. For example, it would be foolish to claim strong implications from one *p*-value being .09 and the other .11, even though the latter does fall outside of the conventional 10% level of significance.[16]

[16] Note that the differences in unobserved heterogeneity mean that, although the significance of the coefficients in each group is informative, the relative *magnitude* of the coefficients across groups is uninformative.

Of course, if a coefficient is (in)significant for both groups, this approach does not provide insight into relative effects. However, the researcher can at least report that, for example, x_1 has a significantly positive impact for both groups.

IMPLICATIONS FOR RESEARCH

The implications for research using the logit or probit model are profound. Conclusions drawn from comparing coefficients across groups while ignoring the possibility of coefficients being confounded with residual variation are meaningless. The simulations in this paper have shown that in the presence of even fairly small differences in residual variation, naïve comparisons of coefficients can indicate differences where none exist, hide differences that do exist, and even show differences in the opposite direction of what actually exists. The substantive importance of this theoretical concern has implications for both gathering data and carrying out statistical testing.

Gathering as complete a set of covariates as possible is more important when using logit or probit than when using linear regression. In the linear case, omitted variables are only significant if they are correlated with included variables. However, in the logit or probit case, any variable that helps explain the outcome variable is useful and should be gathered. The more variation we control for, the less residual variation there is and the less it can vary across groups. We also need a sizable sample to apply the ratio of coefficients technique.

Econometric theory and simulation results suggest that tests interacting coefficients with a dummy variable for group membership in a single equation are particularly misleading. Forcing observations from both groups to have the same residual variation yields coefficients that tell us nothing about how a covariate's impact varies across groups.

Estimating separate equations for each group at least offers the advantage of accurate estimation within each group. However, before attempting to compare coefficients, researchers must test for differences in residual variation. This will require a change in current practice, but the test is simple to run in any statistical package with programming capabilities.

If a difference in residual variation is found, the researcher has several options. Allison's test for determining if at least one coefficient differs between groups is powerful and makes few assumptions. It will lead to more conservative results and may reveal that apparent differences are not actually significant. If it reveals differences, the researcher can then test that a specific coefficient differs. However, this test has a stringent assumption: the other coefficients must be equal across the groups. Since it is not possible to test this, the researcher must judge the probability of this assumption holding on theoretical grounds.

If it is theoretically relevant to compare the relative effects of two covariates across groups, the researcher can compare the ratio of coefficients across groups. This has the advantage of making no assumptions about the residual variation across groups. Offsetting this advantage are two facts. First, an answer in terms of relative effects may not satisfy the theoretical question at hand. Second, even large differences between ratios may not be statistically significant, particularly if one or more terms are poorly estimated. This makes it more difficult, perhaps artificially so, to identify cross-group differences.

Comparing coefficients across groups in logit or probit models requires that the researcher apply careful judgment using his or her understanding of both statistical issues and the underlying phenomenon. Ultimately, however, researchers may simply not be able to conduct some of the comparisons they are accustomed to doing in the linear setting. While this is frustrating, no results are surely superior to spurious results.

Table 1: Apparent coefficient differences in simulated data

	Group 0	Group 1
X1	0.608***	0.376***
	(0.055)	(0.032)
X2	1.379***	0.775***
	(0.104)	(0.051)
Intercept	-0.482***	-0.599***
	(0.127)	(0.103)
Log likelihood	-211.50	-326.14
N	1,000	1,000

Standard errors in parentheses
**** p<0.01;** p<0.05; * p<0.1; two-tailed tests*

The equation

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

was estimated on data generated according to the equation:

$$y_i^* = x_{1i} + 2x_{2i} + \sigma_g \varepsilon_i$$

$$x1, x2 \sim N(0, 4)$$

$$\sigma_0=2; \sigma_1=4$$

Table 2: Differences in residual variation drive apparent differences in estimated coefficients

Scale of residual variation for Group 1 (σ_1)	Mean value of estimated β_2 for Group 1	Number of times β_2 was falsely found to differ across groups (1000 simulations, p=.05)
1.0	2.88	55
1.2	2.42	223
1.4	2.07	681
1.6	1.81	923
1.8	1.61	988
2.0	1.45	997

The equation

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

was estimated on data generated according to the equation:

$$y_i^* = x_{1i} + 2x_{2i} + \sigma_g \varepsilon_i$$

$$x_1, x_2 \sim \mathrm{N}(0, 4)$$

$$\sigma_0=1;\ \sigma_1=1, 1.2, 1.4, 1.6, 1.8, 2.0$$

Table 3: Allison's method accurately detects differences in residual variation and *false* differences in coefficients

Scale of residual variation for Group 1 (σ_1)	Does σ, the scale of residual variation, differ across groups? Number of positive tests out of 1000 datasets (p=.05)		Do any of the coefficients differ across groups? Number of positive tests out of 1000 datasets (p=.05)
	Wald chi-square test	*Log-likelihood test*	*Log-likelihood test*
1.0	70	68	53
1.2	243	163	54
1.4	541	414	46
1.6	805	677	42
1.8	933	880	45

The equation

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

was estimated separately for Group 0 and Group 1 on data generated by

$$y_i^* = x_{1i} + 2x_{2i} + \sigma_g \varepsilon_i$$

$$x1, x2 \sim N(0, 4)$$

$$\sigma_0=1; \sigma_1=1.0, 1.2, 1.4, 1.6, 1.8$$

Table 4: Allison's method also accurately detects *true* differences in coefficients

Additional impact of x_2 on Group 1	**Does σ, the scale of residual variation, differ across groups?** **Number of positive tests out of 1000 datasets (p=.05)**		**Do any of the coefficients differ across groups?** **Number of positive tests out of 1000 datasets (p=.05)**
	Wald chi-square test	*Log-likelihood test*	*Log-likelihood test*
0.0	541	414	46
0.2	383	283	223
0.4	249	187	619
0.6	152	142	893
0.8	103	113	984
1.0	61	90	998

The equation

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (G_i x_2)$$

was estimated separately for Group 0 and Group 1 on data generated by

$$y_i^* = x_{1i} + (2 + \gamma G_i) x_{2i} + \sigma_g \varepsilon_i$$

$$G_i = \begin{cases} 0 \text{ for Group } 0 \\ 1 \text{ for Group } 1 \end{cases}$$

x1, x2 ~ N(0, 4)

σ_0=1; σ_1=1.4

γ ranges from 0 to 1

Table 5: Comparing the ratios of coefficients requires precise estimates of each coefficient

Additional impact of x_2 on Group 1	Does the ratio β_2/β_1 differ across groups? Number of positive tests out of 1000 datasets (p=.05) N=100	Does the ratio β_2/β_1 differ across groups? Number of positive tests out of 1000 datasets (p=.05) N=1,000
0.0	331	41
0.2	353	219
0.4	411	614
0.6	472	886
0.8	518	983
1.0	590	999

The equation

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

was estimated separately for Group 0 and Group 1 on data generated by

$$y_i^* = x_{1i} + (2 + \gamma G_i) x_{2i} + \sigma_g \varepsilon_i$$

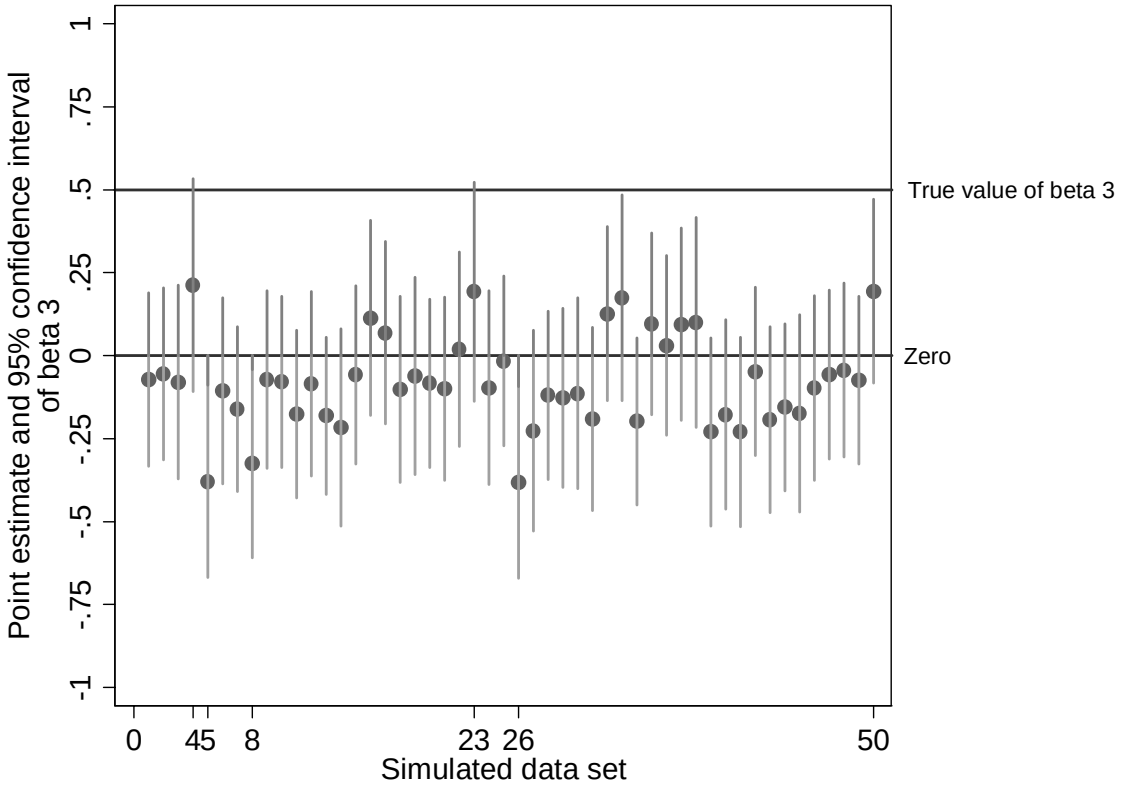
$$G_i = \begin{cases} 0 \text{ for Group } 0 \\ 1 \text{ for Group } 1 \end{cases}$$

$$x1, x2 \sim N(0, 4)$$

$$\sigma_0=1; \sigma_1=1.4$$

$$\gamma \text{ ranges from } 0 \text{ to } 1$$

Figure 1: Estimates using an interaction term are highly misleading



The equation

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (G_i x_2)$$

was estimated on data generated by

$$y_i^* = x_{1i} + 2x_{2i} + .5(G_i x_{2i}) + \sigma_g \varepsilon_i$$

$$G_i = \begin{cases} 0 \text{ for Group } 0 \\ 1 \text{ for Group } 1 \end{cases}$$

$$x1, x2 \sim N(0, 4)$$
$$\sigma_0 = 1; \sigma_1 = 3$$

Reference List

Ai, C.R., and E.C. Norton. 2003. Interaction terms in logit and probit models. *Economics Letters* 80(1): 123-29.

Aiken, L.S., and S.G. West. 1991. *Multiple regression: testing and interpreting interactions*. Thousand Islands, CA: Sage Publications.

Allison, P.D. 1999a. Comparing logit and probit coefficients across groups. *SMR/Sociological Methods & Research* 28(2): 186-208.

Allison, P.D. 1999b. *Logistic regression using the SAS system: theory and application*. Cary, NC: SAS Institute.

Bailey, S.R. 2002. The race construct and public opinion: Understanding Brazilian beliefs about racial inequality and their determinants. *American Journal of Sociology* 108(2): 406-39.

Darnell, A.C. 1994. *A dictionary of econometrics*. Brookfield, VT: E. Elgar.

Davidson, R., and J.G. Mackinnon. 1984. Convenient specification tests for logit and probit models. *Journal of Econometrics* 25(3): 241-62.

Gangl, M. 2004. Welfare states and the scar effects of unemployment: a comparative analysis of the United States and West Germany. *American Journal of Sociology* 109(6): 1319-64.

Greene, W.H. 2000. *Econometric analysis*. 4th ed. Upper Saddle River, N.J. : Prentice Hall.

Gujarati, D. 1988. *Basic econometrics*. 2nd ed. New York: McGraw-Hill.

Liao, T.F. 2002. *Statistical group comparison*. Wiley Series in Probability and Statistics. New York : Wiley-Interscience.

Lincoln, E.J. 2001. *Arthritic Japan : the slow pace of economic reform*. Washington, D.C. : Brookings Institution Press.

Long, J.S. 1997. *Regression models for categorical and limited dependent variables*. Advanced Quantitative Techniques in the Social Sciences. Thousand Oaks, CA: Sage Publications.

Long, J.S., P.D. Allison, and R. McGinnis. 1992. Rank advancement in academic careers: sex differences and the effects of productivity. *American Sociological Review* 58(5): 703-22.

Long, J.S., and M.F. Fox. 1995. Scientific career--universalism and particularism. *Annual Review of Sociology* 21: 45-71.

Maddala, G.S. 1983. *Limited-dependent and qualitative variables in econometrics*. New York: Cambridge University Press.

Orme, C. 1995. On the use of artificial regressions in certain microeconometric models.

Econometric Theory 11(2): 290-305.

Pager, D. 2003. The mark of a criminal record. *American Journal of Sociology* 108(5): 937-75.

Pindyck, R.S., and D.L. Rubinfeld. 1991. *Econometric models and economic forecasts*. 3rd ed. New York: McGraw-Hill.

Quandt, R.E. 1960. Test of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association* 55: 324-30.

Shaver, J.M. 1998. Accounting for endogeneity when assessing strategy performance: does entry mode choice affect FDI survival? *Management Science* 44(4): 571-85.

Simon, R.W., and L.E. Nath. 2004. Gender and emotion in the United States: Do men and women differ in self-reports of feelings and expressive behavior? *American Journal of Sociology* 109(5): 1137-76.

Skeels, C.L., and F. Vella. 1999. A Monte Carlo investigation of the sampling behavior of conditional moment tests in tobit and probit models. *Journal of Econometrics* 92(2): 275-94.

Sorenson, O., and T.E. Stuart. 2001. Syndication networks and the spatial distribution of venture capital investments. *American Journal of Sociology* 106(6): 1546-88.

Train, K. 1986. *Qualitative choice analysis: theory, econometrics, and an application to automobile demand*. Cambridge, MA: MIT Press.

Train, K.E. 1998. Recreation demand models with taste differences over people. *Land Economics* 74(2): 230-240.

Train, K.E. 2004. *Discrete choice methods with simulation*. Cambridge : Cambridge University Press.

Weesie, J. 1999. Seemingly unrelated estimation: an application of the cluster-adjusted sandwich estimator. *Stata Technical Bulletin* 52.

Yatchew, A., and Z. Griliches. 1985. Specification error in probit models. *Review of Economics and Statistics* 67(1): 134-39.

Zeng, Z., and Y. Xie. 2004. Asian-Americans' earnings disadvantage reexamined: the role of place of education. *American Journal of Sociology* 109(5): 1075-108.