臺灣社會學刊 第28期（增刊）
2002年9月 頁231-259
Taiwanese Journal of Sociology
No. 28 (supplement), September 2002

Research Notes

Analyzing Group Inequality: The Problem of Unobserved Outcomes

Tony Tam *

Acknowledgment:
This research was supported by the Research Grants Council of Hong Kong (CUHK 3/93H) and by Academia Sinica through the Organization-Centered Society Project, the Learning 2000 Project, and the Institute of European and American Studies. Direct correspondence to Tony Tam, Academia Sinica, Institute of European and American Studies, Nankang, Taipei, 11529, TAIWAN (email: tam@sinica.edu.tw; tel: (02)2789-9390x249).

* Tony Tam Associate Research Fellow, Academia Sinica
Receipt Date: 2001 / 6 / 1 · Acceptance Date: 2002 / 5 / 23

Abstract

This paper addresses a longstanding problem of analyzing group inequality: when the outcome is continuous but only observed with a binary indicator, the parameters of conventional binary regression models are under-identified. The objectives of this paper are to (1) nontechnically expose the fact that the conventional solution to the identification problem actually results in arbitrary rescaling across models, (2) point out a systematic bias in the estimates of conventional binary regression programs, (3) propose a simple way to implement the Winship-Mare solution, (4) emphasize the parallels between standard and binary regression modeling when analyzing group inequality in a continuous outcome, and (5) present concrete examples of the analysis of group inequality in which a direct comparison of parameters across models is necessary and ignoring under-identification would have serious consequences. Using data from the 1998 General Social Survey of the United States and the 1993 Taiwan Social Change Survey, this paper presents numerical examples from the study of group differentials in occupational achievement (whether a job has high prestige) and school tracks (academic versus vocational schools). The substantive conclusions can be drastically different depending critically on the identifying assumption used.

Key Words: Latent variable, logit, probit, group inequality.

A. Introduction

Group inequality is central to the voluminous literature on stratification. Gender and ethnic inequality, for instance, are two of the most enduring themes of stratification research in the East and West (Grusky 1994). The study of group stratification often entails the measurement of group differences in behavior, achievement, and other outcomes. The outcome is usually interval scale. But often times the outcome is binary. For instance, men and women have distinctly different labor force participation behavior. Children of broken families are more likely than children of intact families to drop out of school. Ethnic groups have different probabilities for attending public universities or becoming medical doctors. All these binary outcomes are significant indicators of the current and future levels of socioeconomic achievement.

When the dependent variable is binary, the method of OLS regression is inappropriate (see, e.g., Long 1997). For the past two decades, logit and probit models are frequently used by sociologists in modeling binary outcomes. For the purpose of this paper, *binary regression models* are restricted to the logit or probit models (Winship and Mare 1983). The widespread availability of computing power and sophisticated-yet-friendly statistical software has made regression analysis of binary data as routine as OLS regression.

However, the technical ease of applying binary regression models has a hidden trap that will be the focus of this paper. Binary regression models often have to solve the *problem of identification* that essentially invalidates any direct comparison of coefficients across models. A naive application of

binary regression models that overlooks the problem can lead to misguided results and conclusions (Cameron and Heckman 1998). Unfortunately, rarely do authors of published articles take heed of the problem. Many users of binary regression models do not seem to be aware of the consequences of the identification problem.

The problem is especially prominent in the analysis of group inequality. The most common analytic task for a study of group inequality is to decompose a group difference into its sources. The task is pervasive in the study of, for instance, gender and racial gaps in wages and educational attainment. But this task requires the direct comparison of parameters across models.[1] When the dependent variable is observed and continuous, the procedure is well understood and routinely implemented in the best quantitative literature of sociology and economics. By contrast, the problem of identification looms large when the continuous outcome variable is unobserved but measured with a discrete indicator - as in logit or probit models. The problem must be resolved before a researcher can carry out the most basic and common analytic task in the study of group inequality.

---

[1] It should be emphasized that a common practice in the literature is to compare the statistical significance levels associated with an estimated effect across different models in an attempt to assess if a group differential can be explained by the inclusion of some explanatory variables. For instance, if the gender wage gap is significant in one model but insignificant after a set of variables are added, the conventional interpretation is that these variables successfully account for the gender gap. Tam (2001) shows that this is a dangerous and uninformed interpretation even when the dependent variable is continuous and observed, and it does not address the need to decompose the gender gap.

Fortunately, a solution has been available since the 1980s (Winship and Mare 1983). The solution can effectively address the identification problem and the analytic requirements of analyzing group inequality. In other words, the solution allows meaningful comparison of parameters across models based on the same data. Unfortunately, the solution has almost never been used. Indeed, the ramifications of the identification problem are poorly understood and, most significant, widely overlooked. While sophisticated methodologists find it too obvious to write a paper on the problem and solution, the social science community is generally ignorant about the problem and its ramifications. There is simply nothing in the literature that pays serious attention to the problem and helps social scientists firmly understand its ramifications.

This paper is designed to fill this void. It is founded on the three basic ideas mentioned above: the problem of identification is a mathematical truth, comparing parameters across models is a relevant task (indeed essential for many substantive contexts), and the Winship-Mare proposal is a sound solution. All three ideas are firmly established knowledge that can withstand any scrutiny. Building on these ideas, (1) I offer an original and nontechnical exposition of the fact that, when there is a latent variable interpretation for a binary regression model, the conventional solution to the identification problem would lead to arbitrary rescaling across models, (2) point out a systematic bias in the estimates of conventional binary regression programs, (3) propose a simple way to implement the Winship-Mare solution, (4) emphasize the parallel between standard and binary regression modeling when analyzing group inequality in a continuous outcome, and (5) present

concrete examples of analyzing group inequality in which the direct comparison of parameters across models is necessary and ignoring under-identification would have serious consequences.

The rest of this paper is organized as follows. Section B provides a non-technical view of the problem of identification and its solution and introduces the first four original contributions of the paper. Section C presents two heuristic examples of the quantitative consequences of ignoring this problem in the study of education and stratification. The first example is concerned with the role education plays in explaining the racial gap in holding high status occupations in the United States. The second example is concerned with the gender and ethnic gaps in getting access to the academic track of schooling in Taiwan. Section D concludes the paper.

B. Modeling Unobserved Outcomes

The statistical modeling of binary data is already well covered in the literature (see, e.g., Winship and Mare 1983; Long 1997). The most common starting point is the postulate of a latent (*unobserved*) continuous variable y* as the theoretical dependent variable of interest. A second postulate is that y* is a function of independent variables, collectively denoted by the matrix X, and β is a vector of effect parameters. That is,

[1] $y^* = X\beta + u$

where u denotes omitted variables and E(u|X)=0. The latent outcome y* is related to an observed binary indicator y with the following straightforward rule for each observation i:

[2] $y_i = 1$ if $0 < y_i^*$

$= 0$ otherwise.

This latent variable approach conceives of binary regression models as an analog of classical regression models for observed continuous dependent variables. The approach is remarkably general and the latent variable encompasses a great variety of concepts.[2] The concepts may be difficult to precisely measure but *in principle* measurable. They may also be inherently *unobservable* but can be indirectly measured with a binary indicator. In any event, the latent variable interpretation is applicable to a wide variety of contexts in which binary regression models are used.

1. Problem of Identification

A distributional assumption for u is necessary for estimating the parameter vector β. The conventional options include the normal and the logistic distribution, resulting in the probit and the logit model, respectively. Although most sociological applications adopt the logit model, the substantive results from the two models should be very similar because the two distributions closely resemble each other (Long 1997:43).[3] As a result, the illustrative empirical examples in the next section will only report results

---

[2] Winship and Mare (1983) discuss a variety of conceptual models that can be appropriately represented within the latent variable framework.

[3] Both the normal and logistic distributions are completely defined by two parameters (mean for location and standard deviation for shape), with the mean (μ_u) and standard deviation (σ_u) of each distribution explicitly entering the pdf and cdf. The parameter estimates under either model are very similar in size if comparable scale assumptions are made (see, e.g., Long 1997). But programming is much simpler for the logit model than for the probit model.

based on the logit model -- the most commonly used model in sociological research.

Just as in the case of standard regression for an observed continuous variable, the parameter vector β depends on the scale of the latent dependent variable y* (Winship and Mare 1983:74; Long 1997:47). Indeed, model [1] will fit the data exactly the same if we multiply y* by any nonzero constant such as 100. The consequence of the change in the dependent variable is simply a new parameter that is a multiple of the original, such as $100\,\beta$. However, unlike the case of standard regression, the dependent variable of model [1] is not observed, its scale is unidentifiable, so is the scale of β. For a probit or logit model, both scales cannot be determined by data. In other words, we must arbitrarily choose a positive variance for u in order to fix the scale of the key parameters of interest without affecting the goodness of fit between a binary regression model and the data. *This is the fundamental problem of scale identification intrinsic to any binary regression model for a continuous latent dependent variable.* The conventional solution is to adopt an identifying assumption that offers the most convenient mathematical form for the purpose of writing program codes:

[3a] $\mu_u = 0$, and $\sigma_u = 1$ if u follows the normal distribution (the probit case),

[3b] $\mu_u = 0$, and $\sigma_u = \pi/\sqrt{3}$ if u follows the logistic distribution (the logit case).

Assumptions [1]-[3] define the results from the standard probit or logit model as implemented in all common statistical packages (such as SPSS, SAS, and STATA).[4]

---

[4] The conventional results may also be interpreted as the parameter estimates for a family of nonlinear probability models that does not postulate the presence of a latent variable y* (see, e.g., Long 1997).

2. Arbitrary Rescaling and Systematic Bias

However, rarely recognized is the fact that the conventional identifying assumption [3] is consequential for substantive interpretation. The assumption gives rise to the problem of incomparable coefficients across models for the same data. As will be explained, the reason is that the assumption results in the arbitrary rescaling of all coefficients across models. Understanding why this is the case is crucial to solving the problem of incomparability.

It is instructive to consider the familiar case of OLS regression for an observed dependent variable y. Two preliminary facts should be noted. First, in a model with an independent variable X_1, part of y's variance would be explained by X_1. When a second model is estimated with the addition of another independent variable X_2, the variance explained necessarily gets higher. That means the variance of the residual u is necessarily reduced. Second, the estimated effect of any independent variable is partly determined by the scale of the dependent variable.[5] If the scale of the same dependent variable changes between two estimation runs of the same data and model, the coefficients of the same independent variable would differ across runs because the metrics of the dependent variable is different for the two runs. We should not compare the coefficients of any variable across models when the dependent variable takes on incompatible metrics in different models.[6]

---

[5] An exception is when the dependent variable y is the logarithm of a positive-valued variable z. In this case, the effect of an independent variable is invariant to any rescaling of z. Only the intercept of a linear model for y will be affected by the rescaling of z.

[6] For instance, for the same data, if you fit a model to the data using USD to code salary

What would the conventional identifying assumption for binary regression models do in the OLS case? If someone imposes the unit variance constraint on the residual of an OLS regression model, the consequence is obvious. As the actual residual variance shrinks with the addition of more variables, the estimated variance (i.e., the scale relative to that of u) of the latent dependent variable would rise. The scale of the dependent variable will increase -- whenever its explained variance increases -- only because by constraint the residual variance is set to unity in all models.

An Important Analogy. A heuristic numerical example from a standard regression model would help fix ideas. Suppose we estimate the gross gender wage gap and then fit a conventional human capital earnings model to see how well human capital differences explain the gender gap. The results (with standard errors in parentheses) are as follows:

[4a] $\text{Lwage} = 5.45 - .15\ \text{Wom} + e_1$

(.15) (.09) $R^2 = .02$

[4b] $\text{Lwage} = 4.61 - .13\ \text{Wom} + .09\ \text{Educ} + .08\ \text{Exp} - .12\ \text{Exp}^2/100 + e_2$

(.06) (.05) (.01) (.01) (.02) $R^2 = .40$

where Lwage stands for the logarithm of observed wages, *Wom* is a dummy variable with 1 for woman, Educ is years of schooling, Exp is years of working experience, and e is the residual. The coefficients of the two models are comparable in the sense that the parameters are defined on the same metrics for the dependent variable. The two models have very different

---

and you fit the same model to the same data but now using NTD to code salary, you are bound to get different coefficients and the coefficients cannot be directly compared.

residual variances (σ^2_e). The unexplained proportion of variances are 98 percent and 60 percent for models [4a] and [4b], respectively. Assuming that the variance of log-wage is 4 in the sample. Then the residual variance for model [4a] is 4*.98=3.92, implying a standard deviation of σ_e=1.98. Similarly, the standard deviation of the residual for model [4b] is 1.55.

Consider now the consequences of fixing the residual variance to be constant and, without loss of generality, fixed at unity across all models. This constraint is equivalent to dividing the residual by σ_e to get a new standardized residual v; σ^2_v is one for all models. In practice, the standardization amounts to multiplying every term on both sides of eq. [4a] and [4b] with constants c_1=1/ σ_e=1/1.98=.51 and c_2=1/ σ_e=1/1.55=.65, respectively. The fit of each model stays the same. The ratios of the coefficients within the same model are identical before and after the multiplication. The parameter estimates under the residual unit variance assumption would be as follows:

[4c] c_1*Lwage = 2.753 - .076 Wom + v_1

[4d] c_2*Lwage = 2.974 - .084 Wom + .058 Educ + .052 Exp - .077 $Exp^2/100$ + v_2

Note that the scale of the dependent variable has to be changed even if those of the independent variables remain the same. Also note that the ratios of the corresponding coefficients in [4a] and [4b] are no longer comparable to those of the corresponding coefficients in [4c] and [4d]. For instance, whereas the gross gender gap in [4a] is *larger* than the one in [4b] after controlling for human capital differences, the gender gap in [4c] is *smaller* than the one in [4d].

In sum, irrespective of the variance assumption imposed on v, the scale

of the dependent variable is artificially altered (i.e., rescaled) from model to model even when the same concept and data are being analyzed in all the models. The scale adjustment does not affect the fit of a model, but it does affect the scale of b - the estimator for β. In fact, the rescaling produces a systematic bias.

Systematic Bias: When the latent variance explained increases, the estimated scale of the dependent variable will be increased, so will the scale of b. This upward bias on the magnitude of b can generate highly misleading results for between-model comparisons. Thus the constant residual variance assumption is a nonsensical restriction: constraining residual variance to be the same across models is useless and would produce misleading results.

Although the heuristic example is set up for an observed dependent variable, the same logic applies to the case of an unobserved outcome with a binary indicator. Whether the dependent variable is observed or not, the constant residual variance assumption produces the same kind of systematic impact on the scale of y* and therefore the scale of b, hence the same kind of incomparability of coefficients would arise. Unfortunately, the same nonsensical assumption is implicit in all binary regression models that assumes [3].

Let us return to the heuristic example and replace Lwage with a latent variable y*. The coefficient of Wom measures the gender gap in y*. Suppose the gender gap is substantially explained by the human capital variables (X). The estimated gender gap should drop after X is added to the model for y*. However, the variance explained of y* would rise as well and the residual variance would decrease. Under the conventional identifying assumption that

fixes the residual variance at unity, all other scales would be measured relative to the residual variance. As the real residual variance declines, the relative scale of all coefficients will rise and the estimated gender gap may become large even though the estimated gender gap might have in fact decreased. The larger the amount of rescaling, the more likely that the reduction in the gender gap may be overcome by the scaling-up of all coefficients.

3. The Winship-Mare Solution

Since the source of the incomparability is an artificial rescaling of the dependent variable across models, the natural solution is to fix the scale of the dependent variable across models, as is automatically done in standard regression models. This is precisely what Winship and Mare (1983:74) suggest when they advocate replacing the sample analog of assumption [3a] or [3b] by setting the variance of y* (σ^2_{y*}) to a constant for all models,[7] such as,

[5] $\sigma^2_{y*} = \sigma^2_{X_b} + \sigma^2_e = 1.$

The computation implied by [5] is considerably simpler than the formula suggested by Winship and Mare (1983:74). The term X_b is simply predicted values of y*. This term is simple to generate by all commercial programs for binary regression.

It is also worth noting that this identifying assumption hinges on the

[7] Since y* is a linear function of Xb and e, the variance of y* is a function of those of Xb and e. Once the variance of e is assumed, the variances of Xb and y* are determined.

postulate of a meaningful latent variable. Without the notion of an underlying y*, assumption [5] would be meaningless. With the postulate of an underlying y*, however, *the constancy of variance should apply to y*, not to the residual term e.* After all, the same dependent variable in the same data is the focus of all the models to be compared. No valid solution should ignore or violate this basic fact. Indeed, the Winship-Mare solution rests on this fact. Once made explicit, it is simply hard to think of any alternative as natural as this one.

Granted assumption [5], then, we can compare coefficients across different models for the same data and dependent variable, just as we can in the case of OLS regression. This solution works for binary regression models based on the normal or logistic distribution. The main precaution is to ensure that different models are indeed estimated on the same (or virtually the same) sample. This is not always the case as models with different independent variables can have very different missing data problems.[8]

An additional practical advantage is that *the size of all effect parameters can be interpreted in terms of the standard deviation of y**. An effect of 0.5 for any independent variable means that a unit change in the independent variable would lead to half a standard deviation rise in the expectation of y* -- the same interpretation for all independent variables and across all models. Yet another advantage is the availability of *a well-defined measure of R-squared for y**. The variance of the latent error term is equivalent to the

[8] It is therefore a good practice to include only cases with valid data for all models to be compared.

variance of y* that is unexplained. To distinguish it from the conventional R-squared, I call it the *Latent R-squared* to signify the fact that the new R-squared applies to the latent dependent variable (y*).

In sum, the seemingly benign technical problem of identification is consequential in practice. The problem is important because ignoring it could seriously distort substantive findings and conclusions. Nonetheless, few researchers take heed of the consequence. Most users are apparently ignorant of the fact that the coefficients estimated with the conventional identifying assumption are not directly comparable across models. A good way to drive home the lesson is to provide numerical examples. Although Winship and Mare (1983) propose the solution in [5], they do not demonstrate that potentially misleading results could emerge under the conventional identifying assumption. Section C will demonstrate the consequences of the conventional identifying assumption in the context of analyzing group inequality.

4. Parameters of Interest

The heuristic numerical example of [4] also underscores the fact that binary regression modeling for unobserved continuous dependent variable (y*) is fundamentally analogous to standard linear modeling for observed continuous dependent variable (y). When there is a latent variable interpretation for a binary regression model, it is a perfectly legitimate extension (indeed the most natural analog) of the conventional logic of linear models to regard the linear parameters (β) as the parameters of interest. In other words, group inequality in a continuous outcome is best summarized by

the parameter of group difference in the linear model for the continuous outcome. As will be demonstrated in section C, sources of the group inequality can be specified by comparing the parameter of group difference across models. Whenever the latent variable interpretation holds, there is simply no compelling reason to deviate from the well tested and well established procedure of analyzing group inequality when the dependent variable is observed and continuous. Whether the continuous dependent variable is observed (y) or unobserved (y*), the same logic of analyzing group inequality should be applied.

In this light, the common practice of turning attention to group differences in the probability of an outcome simply evades (rather than legitimately substitutes for) the well established procedure of analyzing group differences based on linear parameters (β). It is important to recognize that the effect of an independent variable on the probability of an outcome is a nonlinear function of *all* independent variables. In stark contrast to the simple effect parameters of a linear model, there is no single probability effect for any independent variable. In lieu of a single effect parameter, the probability effect of an independent variable is generally an infinite set and the effect is contingent on the level of all other independent variables (Long 1997:63-64). Practitioners do come up with a variety of conventions to sidestep this embarrassing complexity. However, the conventions never change the fact that we will obtain different decomposition of group inequality depending on the particular probability effect picked.[9]

---

[9] For instance, Cameron and Heckman (1998:281-284) examine in detail the estimation

This lack of invariance highlights the fact that probability effects in binary regression are not structural parameters for group inequality. One of the missions of science is to search for structural explanatory mechanisms. The usual hallmark of a structural mechanism is that the mechanism is well captured by a small number of parameters and yet able to generate complex outcomes. The relationship between a structural model to a set of complex outcomes is analogous to that between a multiple regression model and a large number of bivariate correlations among many variables. A binary regression model for an unobserved continuous dependent variable potentially represents a structural mechanism. The linear parameters are then structural parameters. However, if a researcher translates a binary regression model into the corresponding set of predicted probability effects, the result would generally be an infinite set of probability effects. These complex predictions are a big step backward from the parsimony of structural parameters.

There is an alternative way of understanding the fundamental difficulties with looking at probability effects. Consider estimating a linear structural model for an *observed* continuous dependent variable. We can always specify a way to dichotomize the dependent variable and derive effects on the probability of the binary outcome for each independent variable. But

---

results of educational transitions in which probability effects prove to be an unreliable and arbitrary basis for comparing effects across contexts. While methodologists may be content with a mechanical interpretation of a model, social scientists should be concerned with structural parameters that reflect critical sources of differences. The substantive problem of analyzing sources of group inequality precisely calls for such a principle.

everyone would agree that it makes no sense to report these probability effects instead of the linear effect parameters as a way to characterize the structural mechanism. If the mechanism has a simple linear structure for the continuous dependent variable, there is every reason to regard the linear parameters as structural parameters. Especially when the substantive question is about the sources of structural group inequality, the linear parameter of group difference should be the primary reference for analyzing the sources of group inequality. Similarly for an unobserved continuous dependent variable, it does not make sense to substitute probability effects for linear effect parameters. Whether the continuous dependent variable is observed or unobserved, the same logic of analyzing group inequality should be applied. To drive home this message, the next section will demonstrate the logic of analyzing group inequality with two numerical examples.

C. Analyzing Group Inequality

1. High Status Occupations

The first example is a straightforward illustration of how the numerical results can differ under the conventional and Winship-Mare identifying assumptions. The analysis draws on a study of American racial differences in occupational prestige and how the racial gap may be related to educational attainment and father's education. For the purpose of illustrating the problem of comparing across binary regression models, I will analyze the racial gap in the propensity to attain a "high status" job -- a binary outcome variable that defines a high status job as one with occupational prestige higher than 50.

When restricting attention to adults of at least 24 years old, the 1998 General Social Survey (GSS) shows that only 33% of the cases attained high status.[10] Thus these high status jobs in the United States are the top one-third in terms of occupational prestige.

Table 1. Racial Gap of Access to High Status Occupations in the United States: GSS 1998 (N=1,841)

	A. Conventional Estimates[1]		B. Rescaled Estimates[2]	
Independent Variables	(1)	(2)	(3)	(4)
Nonwhite	-.613*	-.449*	-.335*	-.209*
	(.148)	(.162)	(.081)	(.075)
Years of Schooling		.385*		.180*
		(.025)		(.012)
Father's Education		.004		.002
		(.014)		(.007)
Constant	-.486	-6.002	-.266	-2.799
Latent-R^2			.014	.284

Standard errors in parentheses.
[1] Panel A: Logit estimates with conventional identifying assumption.
[2] Panel B: Logit estimates with Winship-Mare identifying assumption.
* Significant at .05 level.

To assess the sources of the racial gap in attaining high status jobs, I will have to compare the estimated racial gap across two models.[11] The first model estimates the gross racial gap and the second model estimates the

[10] I use the prestige score associated with the occupational code of a respondent based on the 1980 Census Occupational Classification for the current or the last held job. The prestige scores in the public release file of the 1998 GSS are appended to the data by the National Opinion Research Center.

[11] All analyses were conducted with Stata 6.0 (StataCorp 1999).

conditional (net or residual) racial gap after controlling for a respondent's years of education and father's years of education. Two sets of results are presented in table 1. Panel A presents the white-nonwhite gap in attaining a high status job, with and without controlling for the educational background variables. The conventional approach is to directly compare the coefficients of Nonwhite across columns (1) and (2) and interpret the reduction in the gap as the extent of the gross gap attributable to racial differences in educational attainment and parental education. This procedure suggests that 27% of the gross gap can be explained.[12] Panel B estimates the same models with the same data but imposes the Winship-Mare identifying assumption. The estimates of columns (3) and (4) are parallel to those of columns (1) and (2), respectively. What about the gross gap attributable to racial differences in educational attainment and parental education? Comparing the coefficients of Nonwhite across columns (3) and (4) suggests that 38% of the gross gap has been explained.[13] The fraction of the racial gap explained is substantially higher (*38% instead of 27%*) than what the conventional procedure suggests.

We can find a hint of the source of the discrepancy by comparing the Latent R-squared across models - that is, the portion of the variance of the latent dependent variable that has been explained. The Latent-R^2 are 1% and 28% for the models in columns (3) and (4), respectively. With the Latent-R^2 so vastly different, the coefficients under the conventional and Winship-Mare assumptions are bound to be quite different. In addition, the impact of the

---

[12] 100% ×(.6125883 - .4491311)/.6125883

[13] 100% ×(.3353079 - .2094838)/.3353079

arbitrary rescaling due to the conventional identifying assumption is likely to be large as well. But the discrepancies are not necessarily just a matter of degree. The discrepancies sometimes can lead to qualitatively different conclusions, hence producing misleading results with dire consequences. This will be illustrated in the next example.

2. School Tracks

The substantive dependent variable of the second analysis is school track attainment. In Taiwan, educational inequality has at least two dimensions: quantity and track. Quantity refers to the years of schooling whereas track refers to the membership in the academic (senior high, university) or vocational (vocational high, vocational college) systems. For a long time each track entailed a fairly rigid path of educational development. Switching track was rare. The empirical analysis here will focus on measuring the group inequality of academic track attainment, i.e., whether the track is academic or vocational. Only graduates of senior high schools, universities, or graduate schools would be coded one for the binary dependent variable. Graduates of vocational high schools, vocational colleges, junior high schools or less would be coded as zero.

Pooling data across two subsamples (each responded to a different questionnaire with overlapping items) of the 1993 Taiwan Social Change Survey (TSCS), we can estimate educational inequality between women and men, and across ethnic groups in Taiwan. The number of valid cases is 3,898, which turned out to be not large enough for precise estimation. I therefore derive a modified data set from this sample by expanding the same size ten-

fold to 38,980. By design, this partially simulated data set retains all the statistical properties and relationships among the variables in the original sample.

Table 2. Gender and Ethnic Gaps in Academic Track Attainment in Taiwan: Modified TSCS 1993 (N=38,980)

	A. Conventional Estimates[1]		B. Rescaled Estimates[2]	
Independent Variables	(1)	(2)	(3)	(4)
Women	-.398*	-.463*	-.214*	-.240 *
	(.026)	(.026)	(.014)	(.013)
Mainlanders	1.025*	.977*	.552*	.505 *
	(.034)	(.034)	(.018)	(.018)
Aborigines	-.315	-.378*	-.169	-.196 *
	(.171)	(.172)	(.092)	(.089)
Age		-.015		-.008
		(.009)		(.005)
Age-squared/100		-.032*		-.017*
		(.010)		(.005)
Constant	-1.333	-.006	-.718	-.003
Latent-R^2		.077		.357

Standard errors in parentheses.
[1] Panel A: Logit model with the conventional identifying assumption.
[2] Panel B: Logit model with the Winship-Mare identifying assumption.
* Significant at .05 level.

Table 2 presents two panels of estimates on the gender and ethnic inequality of academic track attainment in Taiwan, without and with control for age cohort. From preliminary analysis, I find that Fukien and Hakka Taiwanese are indistinguishable in the analysis. So the estimates in table 2 consider Fukien and Hakka Taiwanese together as a single reference group. Since aborigines and other ethnic groups are a tiny minority, the empirical focus would be on Mainlanders versus Taiwanese. The estimates of panel A

are based on the conventional logit estimates, with identifying assumption imposed on the variance of the latent error term. After controlling for nonlinear age cohort effects, the women disadvantage of attaining the academic track *widens* by over 16% (from -.398 to -.463) and the Mainlander advantage over Taiwanese *narrows* by less than 5% (from 1.025 to .977). In other words, taking into account differences in the age cohort composition of different groups, gender inequality is substantially larger than the gross gap suggests whereas ethnic inequality is slightly smaller than the gross gap shows.

Panel B of table 2 presents rescaled estimates based on the Winship-Mare identifying assumption. The contrasts with panel A are interesting. After controlling for age cohort, the gender gap of panel B widens by about 12% (from -.214 to -.240) while the Mainlander advantage drops by more than 8% (from .552 to .505). Thus controlling for age cohort composition has a smaller impact (*12% instead of 16%*) on estimated gender inequality but a larger impact (*over 8% instead of less than 5%*) on estimated ethnic inequality than the conventional logit estimates would suggest.

It is well known that educational expansion in Taiwan is a recent phenomenon. Younger cohorts benefit more from the expansion than do older cohorts. But the expansion is biased toward the vocational track in the decades before the 90s. Controlling for age cohort makes a difference for both the gender and ethnic gaps, but in opposite ways. Because women in the sample happen to be younger than men are, and women face a disadvantage, controlling for age widens the gap. Whereas Mainlanders are younger than Taiwanese because most Mainlanders came to Taiwan after World War II,

controlling for age can explain away more of the advantage for Mainlanders. Thus a naive reading of the conventional outputs from logit models may lead to a substantively different story of how age cohort accounts for the gender and ethnic gaps in academic track attainment in Taiwan.

D. Conclusion

The problem of unobserved achievement outcomes is basically a problem of identification inherent to conventional binary regression models -- whether it is the probit or logit model. All available statistical packages adopt the same identifying assumption. However, the conventional identifying assumption is anything but benign. In fact, the assumption results in the awkward situation that the scale of the latent dependent variable is artificially altered from model to model, even when the sample is identical across models. The consequence is the incomparability of coefficients across models, indeed a systematic bias on the reported estimates, and therefore a fatal trap for researchers studying group inequality through model comparison. Ignoring the problem would be like insisting on comparing apples with oranges. In order not to jeopardize the validity of any substantive inference, the problem of incomparability must be addressed rather than ignored.

Fortunately, a sound solution is available. The proper solution is to do what Winship and Mare (1983) advocate, namely, fix the scale for the latent dependent variable as we always do with OLS regression. This paper has demonstrated through two empirical examples the numerical consequences of ignoring the incomparability. The findings are illustrative of how misleading

a naive reading of the conventional outputs can be. Comparing coefficients estimated under the conventional identifying assumption is vulnerable to the systematic bias stemming from the rescaling of the latent dependent variable. Before any comparison of coefficients across models, it is imperative that all estimation results be adjusted according to the Winship-Mare identifying assumption.

It should also be emphasized that the problem of incomparability are not confined to the substantive context of analyzing group inequality. Most quantitative studies in sociology and other social sciences would involve comparing coefficients across models -- much more often than many researchers think (Tam 2001). The specification of multivariate relationships among a set of variables usually entails the estimation of multiple multivariate models and comparing the size of coefficients across different models. This analytic task is germane to virtually any quantitative theory building in all substantive areas of research. Therefore it is only prudent that applications of conventional binary regression models routinely adopt the Winship-Mare identifying assumption in estimating effect parameters. Currently no statistical package would offer this option. The Winship-Mare assumption (condition [5]) must be imposed manually. But this cannot be an excuse for ignoring the problem. After all, this paper has demonstrated that the requisite rescaling is not difficult to do and involves only simple arithmetic.

Tony Tam is associate research fellow at the Institute of European and American Studies, Academia Sinica. His main research interests are stratification and labor market, education, organizational and economic sociology, and statistical methodology. He is currently the Principal Investigator of "Learning 2000: Educational Institutions and the Creation of Human Capital" -- a collaborative project with six basic research subprojects on education. Among the subprojects is a strategic initiative to conduct the first national panel survey of about 40,000 Taiwanese high school and five-year specialized school students, their parents, teachers, and schools.

BIBLIOGRAPHY

Cameron, Stephen V. and James J. Heckman, 1998, "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males." ***Journal of Political Economy*** 106:262-333.

Grusky, David B., ed., 1994, ***Social Stratification: Class, Race, and Gender in Sociological Perspective.*** Boulder, CO: Westview Press.

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

StataCorp, 1999, ***Stata Statistical Software: Release 6.0***. College Station, TX: Stata Corporation.

Tam, Tony, 2001, "Three Common Myths in Quantitative Social Research." ***Taiwanese Journal of Sociology*** 26:253-284.

Winship, Christopher and Robert D. Mare, 1983, "Structural Equations and Path Analysis for Discrete Data." ***American Journal of Sociology*** 89:54-110.

群體差異的分析：隱性變項之難題

譚康榮

中文摘要

本文呼籲學界注意長久存在於群體不平等研究中的一個困境：群體不平等的研究問題往往需要做不同模型間參數的比較。然而，當研究的依變項不能直接測量，能觀察到的是二元的指標，此時模型間的參數是不能直接比較的。原因是，傳統二元迴歸模型的參數是不能識別的（under-identified），能估計的是參數的相對值，而非絕對值。針對前述問題，本文的重點有五：（1）指出傳統上用以解決識別不定性問題的方法，實導致了模型間量度的任意變換；（2）指出一般統計軟體所提供的二元迴歸估計值，呈現有系統偏差；（3）提出簡化的程序以執行Winship-Mare所提出的解決方法；（4）當分析群體不平等的來源時，無論依變項是可觀察與否，只要依變項是連續變項，詮釋標準迴歸模型和二元迴歸模型參數的邏輯應該都是一樣的；和（5）利用1998美國General Social Survey和1993台灣社會變遷的資料，以職業取得（是否聲望高的職業）和學校分流（高中和高職）相關研究的數據做爲例子，具體呈現當實質的研究課題的確需要做模型間的結構參數的比較時，如果忽略了識別不定性問題，對研究結果所造成的嚴重偏誤。

關鍵字：隱性變項、邏輯迴歸、波比迴歸、社群差異

作者簡介

Tony Tam（譚康榮）是中央研究院歐美研究所副研究員。主要研究興趣爲社會階層與勞動市場、教育、組織與經濟社會學，及計量方法。他目前是「學習2000：教育制度與人力資本之創造」計劃主持人。這個大型教育基礎研究共有六個子計劃，其中的「教育長期追蹤資料庫」收集將近四萬名台灣國中、高中／職及五專學生、他們的家長、老師和學校的資料。