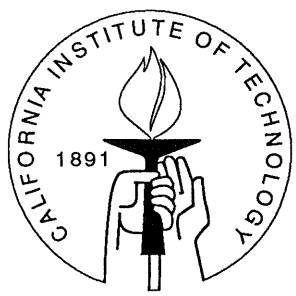
DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91125

Regression Options for Historians: Choosing Among OLS, Tobit, Logit, and Probit Models

Douglas Flamming



HUMANITIES WORKING PAPER 152

September 1992

Douglas Flamming

Regression Options for Historians: Choosing Among OLS, Tobit, Logit, and Probit Models

Only a generation ago, it was almost unheard of for historians to use multivariate regression analyses to substantiate their arguments. And when historians did begin to utilize regression models in the early 1970s, the suitability of such methods for historical research was a matter of considerable debate.[1] But today, although protests occasionally resurface[2], regression tables can be found in mainstream journals and well-received books, on topics ranging from the impact of the Counter Reformation in France to the dynamics of kinship in colonial New England to the nature of the Ku Klux Klan in Indiana.[3]

Now that multiple regression is regularly used and generally accepted within the profession, historians need to take a calculated second look at the regression options at their disposal. In almost every instance, historians utilize the standard "ordinary least squares" regression technique (OLS).[4] As a widely discussed and readily accessible method of analysis, OLS has been and continues to be the most popular regression technique employed by social scientists.[5] It gives dependable estimates under a wide variety of empirical circumstances. But many historians remain unaware of the other regression options commonly utilized by econometricians--tobit, logit and probit. Almost a decade ago, historian Philip T. Hoffman pioneered the use of tobit and probit in historical research, but studies utilizing those techniques are still few and far between. Such options are nonetheless of critical importance to those who regress, because the data that historians have to work with may often be inappropriate for an OLS model and is often well suited for one of these alternative regression strategies. And since logit, probit, and tobit programs have recently become readily available for personal computer users, there is no reason for historians to use OLS when the data calls for another method.[6]

This essay provides, first, a brief and largely nontechnical discussion of the different types of regression analysis available to historians and when they should be used. It then offers two case studies to demonstrate the importance of, and sometimes the difficulty of, choosing the proper regression technique. The first case is an analysis of women's wages in the southern textile industry in the early twentieth century, and the second is an investigation into why mill workers in a Georgia textile mill divided over a prolonged strike in 1939. In both studies, the proper approach would seem, intuitively, to be a standard OLS regression model; but, as shall be

demonstrated below, the data in both cases proved problematic for OLS and required the use of alternative forms of regression analysis.

The four types of regression discussed here--OLS, tobit, logit, and probit--are technically distinct, but they share some common assumptions and characteristics. Anyone familiar with the logistics of OLS regressions will have no difficulty using alternative forms. In each method, a single dependent variable y is said to be a function of a collection of independent variables. The independent variables are also called explanatory variables because they "explain" changes in y. All four methods produce numeric scores--called coefficients--for each independent variable. These coefficients indicate the relative effect each explanatory variable has on y--independent of the other independent variables. The principal goal of any regression method is to determine how strongly related each explanatory variable is to the dependent variable when the effect of the other independent variables is taken into consideration. There are nonetheless important differences in interpreting the coefficients when using the various methods (to be discussed below). Certain assumptions are common to all four methods, including the critical injunction that multicollinearity among the independent variables must be avoided.[7] In each type of model, it is permissible to transform the independent variables, when necessary, by logarithmic calculations or by weighting. It is also acceptable in each method to include dichotomous explanatory variables (commonly called "dummy" variables), the values of which (usually 0 and 1) simply indicate the presence or absence of some characteristic. Finally, it is critically important that the statistical viability of the coefficients is, in each method, determined by similar tests of statistical significance, such as t-scores.

Making the right decision about which type of regression to use hinges upon a careful consideration of one's dependent variable. The dependent variable is represented by y in the standard OLS equation:

$$y = a_0 + b_1x_1 + \ldots b_nx_n + u$$

where x_1 though x_n are independent "explanatory" variables, b_1 through b_n are unknown parameters, a_0 is a constant term, and u is an error term. All OLS equations require that y represent an interval-scale of values, such as income (in dollars), temperature (in degrees), or age (in years), and the data is assumed to be taken from a random sample. Suppose we had a sample of one-hundred workers employed by a hypothetical company named Laborco, and we wanted to know which type of workers earned the highest wages--commonly called a wage-earnings functions analysis. Our dependent variable would be weekly wages, and our independent variables would be a collection of individual characteristics, such as age, gender, education and

race. So our proposed y--the workers' wage rates--would range between a minimum value and a maximum value, falling along an interval scale. This type of dependent variable is appropriate for an OLS regression, even if the one-hundred data points for "earnings" were not exactly normally distributed.[8]

When the dependent variable y is statistically "limited" in some way, OLS equations become untrustworthy and an alternative must be found. *Limited-dependent variables* come in several forms. One is the "truncated" variable. A truncated y is one whose distribution is "cut off" at some arbitrary value owing to the nature of the sample taken. Taking once again our hypothetical example of Laborco, we might want to determine the earnings functions of wage-earning employees who earned more than n dollars an hour. We would therefore exclude from the sample any workers who earned an hourly wage less than n. Our sample would not be random; it would not represent the poorest-paid employees at Laborco who fell below the minimum level of inclusion. The dependent variable "Hourly Wage" would therefore be considered limited because of the deliberate truncation of the sample.

A "censored" dependent variable is one in which observational differentiations are masked by the nature of the data. Economists encounter this problem primarily in consumer expenditure analyses. The classic example involves annual household expenditures for automobiles. Any attempt to explain expenditures y by a collection of household characteristics is marred by the censorship inherent in the values for y. The idea is that any household will doubtless have some desire for an additional vehicle, but the family will not purchase another until that desire exceeds some unspecified threshold--the point at which they simply have to have it and will pay the price for it. For any given year we will have a large group of households with zero expenditures for an additional automobile, but interpreting those zeros is no simple matter. One family scoring zero might actually be very near its threshold and would be on the verge of buying a new car; upon doing so, its measured expenditures would suddenly be quite large. Another family might have no intention of ever buying another car. Can the "zero" for that family be compared to the zero for the family who is almost ready to make a purchase? A good deal of statistical knowledge is "censored" by the nature of the data. The dependent variable for automobile purchases is therefore said to be censored at the threshold, since we do not see any purchases at prices between zero and the threshold.[9]

For historians, the following example of censoring may be more relevant. Suppose a group of social historians decided to undertake a statistical analysis of property distribution for selected counties in a single state throughout most of the nineteenth century. They wanted to determine what personal characteristics were most strongly related to real wealth. County tax rolls provided

the necessary data on property holdings, and samples on individual taxable property were taken for each census year from 1850 to 1880. For each household head listed in the census, the researchers obtained a value for property holdings. But as it turned out, the tax collectors in that state during those years were required by law to ignore property valued at less than $50. As a result, some number of individuals listed in the data base as propertyless were actually petty property holders. So if these historians used taxable property as their y in a regression model, their dependent variable would be censored.

There is no set rule for determining the severity of a censoring problem. The dangers inherent in a censored y depend less upon principle than upon the specific data being used. Suppose only a small minority of individuals on the tax rolls owned property valued between zero and $50. The data for y would technically be censored, but the impact of this cut off point on the distribution of y would be trivial. In such a case, OLS would still be an acceptable method. But historians using OLS should carefully consider whether their dependent variable presents a legitimate censorship problem. Whether the artificial cut off is at the maximum or minimum end of the scale is of no consequence; what matters is that, in a substantial number of cases, the dependent variable represents "censored" values rather than "actual" values.[10]

Technically, censoring and truncation are separate problems, but in the econometrics literature the terms are often used synonymously, largely because both types of data pose the same problem for OLS equations. The problem may be stated simply: when using a truncated or censored dependent variable, OLS will almost always underestimate the regression coefficients. The reason for this is illustrated in Figure 1.[11]

[Figure 1 about here]

When using a truncated or censored y, the correct alternative to OLS is tobit analysis. Tobit analysis accommodates and accurately accounts for numerically censored dependent variables. When using OLS, we think of the dependent variable as being a linear function of our independent variables, and we say the same of a tobit model, although tobit coefficients are derived differently. Whereas OLS gives us an exact solution using a single equation, tobit uses an iterative procedure to produce a very close approximation. Tobit has an added advantage for historians, for by accurately dealing with truncated and censored dependent variables, tobit can be used when the sample under study (which may be as good as the historical record allows) is not representative of the population.[12]

Another common form of limited dependent variable is one that has only categorical values. Sociologists and political scientists confront this situation when evaluating survey responses such as "agree somewhat," "agree," or "strongly agree." It may be assumed that these

three responses represent some logical hierarchy of values from lowest to highest, but they are not interval-level observations, since there is no standardized distance between one response and another. More often, categorical dependent variables have only two possible values. For example, suppose our earnings data for Laborco only indicated whether or not an employee took home more than $10,000 per year. Our dependent variable would be dichotomous, or "discrete." Individual workers either earned more than ten grand, or they did not; y therefore operates as a conventional dummy variable. This binary data would offer a rough gauge of each worker's earnings power and would therefore allows us to ask similar questions about earnings functions.

Categorical dependent variables are inappropriate for OLS models. Assume that a dichotomous dependent variable is being used where y=1 if a worker earned more than $10,000; and y=0 otherwise. One of the important assumptions of OLS models is that the error term will have constant variance across all of the observations (homoskedasticity), an assumption violated by a dichotomous dependent variable. To understand this violation, consider the models illustrated in Figure 2. The first illustration, Figure 2a, is a standard bivariate OLS model. The observations (dots) appear loosely bunched in a nice, equal-sized band on both sides of the regression line, indicating that homoskedasticity exist. Suppose, though, that the same regression line was surrounded by tightly paced observations at the lower end and widely scattered observations at the high end. In such a case, a condition of heteroskedasticity would be evident, and a principal assumption of OLS would be violated. As a result, the OLS coefficients might still be accurate, but the significance scores would be untrustworthy. Consider, then, how an OLS regression line would fall if y were a dichotomous variable (see Figure 2b). By default, a heteroskedastic condition exists, since there is no possible way for the data points to be clustered in an even band along the regression line.

[Figure 2 about here]

When y is a categorical variable, a logit or probit analysis should be used. Like tobit models, logit and probit work by means of iterative algorithms to produce coefficients that are tolerably close estimates of the observed relationships between the dependent variable and each independent variable. Although logit and probit are technically distinct methods, they will generally produce very similar results. Unlike OLS models, they do not assume a linear relationship between y and the independent variables. In Figure 2 the standard bivariate OLS model (see 2a) is compared with the S-shaped cumulative normal distribution that is assumed by both logit and probit (see 2b). Figure 2b also reveals another basic problem with using a dichotomous y in an OLS equation: the model will produce inaccurate predictions, including scores that are greater than 1 and less than 0, a logically impossible result. In this example, the

OLS line nearly parallels the S-curve at some points, but at the extremes, the OLS line will always veer away (as it does here) from the more accurate S-curve.[13]

One cannot directly compare logit or probit coefficients with OLS estimates. OLS and tobit models allow us to say that one-unit of change in the independent variable corresponds to *n* units change in *y*. And since the relationship is assumed to be linear, we may say that the extent of change was constant across cases in the sample. So, for a normal OLS wage-earnings functions analysis at Laborco (or a tobit), a coefficient of 22.5 for the variable AGE would indicate that, on average, a one year increase in an employee's age would mean an extra $22.50 in her weekly paycheck. However, the estimates produced by logit and probit, based on the S-curve distribution, obviously do not have the same predictive quality. Cases falling on different parts of the curve would experience dissimilar magnitudes of change. This problem of comparison presents more of a dilemma for public policy studies than for historians, who are often less concerned with predictive outcomes than with accurately identifying the collection of factors that had a significant effect on *y*. Historians, that is, are usually more interested in identifying the multiple dynamics of change than the precise magnitude of change induced by a particular variable. As a result, scores indicating statistical significance are usually more important to us than the size of our coefficients. Fortunately, then, OLS, tobit, logit, and probit equations all produce standard errors, which allow for the computation of t-scores and other significance tests.[14]

Determining the best regression technique to use thus appears to be a rather straightforward matter. The OLS method should be used when the dependent variable *y* is a continuous interval-scale variable with no unusual value limitations. When the values of *y* are censured or truncated, a tobit model should be employed. If *y* is a categorical variable, logit or probit analysis is the right choice. Historians who know these basic guidelines will likely use the right method for the data at hand. But a theoretical knowledge of when to use logit instead of OLS is only the beginning. Historical data, being decidedly unwieldy, often limits what we can measure. The available numbers may also force us to rely upon a dependent variable whose "type" is difficult to categorize. As a result, choosing the proper regression strategy is not always easily determined, as the following two case studies demonstrate.

Consider the regression quandary I recently encountered when trying to understand the statistical relationship between industrialization and women's liberation. One of the central questions in the social history of the western world is the extent to which industrial development fostered greater individual autonomy for women in society. One argument has been that economic modernization helped liberate working women by loosening their ties to the patriarchal

family farm and workshop and putting them in a situation in which they could earn wages independently of their parents. Some recent works in labor history and women's studies have attacked this argument as fundamentally misleading on several grounds. One point in rebuttal is that working class women--particularly young, single "working girls"--seldom kept their wages. Instead they turned their earnings over to parents, in accordance with common notions of the family economy, thereby cementing traditional female family roles within the industrializing process. The impressionistic evidence for this argument is strong, but few quantitative analyses have been done on the issue of keeping wages.[15]

During the first decade of the twentieth century, the federal government compiled data pertaining to working girls and their wages in the American textile industry. With this data I sought to analyze empirically the issue of female economic independence. My study focused on the "cotton mill girls" of Georgia, a state representative of the textile South, in which cultural norms of fatherly dominance, familial loyalty, and female subservience were deeply entrenched. The data for Georgia therefore offered a useful case study for analyzing the meaning of textile wages for working-class girls. The data consisted of 276 single women workers investigated by government agents in thirty-one cotton mills of varying sizes and locations throughout the state. The surveying began in late 1907 and continued through the spring of 1908. Only single women over sixteen years of age living at home were included in this particular sample. For each woman, the government gathered data on a variety of individual characteristics, such as the number of years she had been working, whether she was literate, and how much she earned during the year. Data was also gathered about her family: the number siblings in her household, the occupation of her father, and the total income of her family that year. The data is, to say the least, a rich source for analyzing the dynamics of women's factory work and family relationships.[16]

A statistical profile taken from the data reveals the following. The average age of the women was 19.6 years, but their relative youth did not mean they were novices in the factory. They had been at work, on average, for 6.4 years. Three-quarters of the working girls were literate, but the ability to read and write did not have any relation to better jobs. Most worked on lower-paying jobs in the mill; not quite one-fifth were weavers, the highest paying job for females in the cotton mills. The women earned an average yearly income of $243. They came from large families, with an average of 2.6 siblings sixteen years or older, and an average of 2.4 siblings under sixteen. Slightly more than two-thirds (71 percent) had fathers living at home (some of whom did not work for wages), and only about 10 percent had mothers who worked for wages. Among those fathers who were gainfully employed, only 41 percent were textile workers. The nontextile fathers were a diverse lot, ranging from farmers to common laborers to petty proprietors. The women in the sample came from households whose average family income for

the year was $1,088. On average, then, each working girl's earnings represented 22 percent of total household income, a figure that clearly indicates how important her earnings were for the family economy.

From the data I was able to determine the amount of annual textile earnings personally retained by each cotton mill girl, that is, the amount of money she got to keep after contributing her share to the family coffer. The data for each household included the number of children over sixteen years of age and the total amount of earnings kept by those children. To calculate the amount kept by each working girl--the variable SHARE--I assumed that each child over sixteen kept an equal share of the total amount kept. Then I divided the total amount kept by the number of siblings over sixteen, the result of which represented the estimated share, in dollars, that the working girl was able to keep. For some girls, SHARE represents an estimate rather than the actual amount retained. Values are "real" for women who kept nothing (the vast majority) and for those who were the only child over sixteen in the household (this latter group being decidedly few). For those who kept some wages and had one or more siblings over sixteen (27 percent of the sample) SHARE is an estimate.

One possible problem with the estimated values of SHARE is that they do not take the gender of siblings into account (the data do not indicate the gender of siblings). It might be argued that male siblings were more likely to be wage-keepers than female siblings. If so, then my estimates for girls from households in which male children predominated are inflated. On the other hand, there were doubtless households in which older brothers were farm workers and sisters were housekeepers, neither of whom would have brought in steady wages from which some earnings could be retained. In such a case, a cotton mill girl would likely have kept a greater share of her wages than my estimate indicates. Fortunately, then, the two potential biases push in opposite directions, so, on average, they may well cancel each other out. And since they comprise less than a third of the sample, the variable SHARE may, I think, be used with confidence.

Working girls who got to keep any of their wages at all were the exception. Table 1 indicates the distribution of the variable SHARE. Obviously, single females working in the cotton mills of Georgia in 1907-1908 seldom enjoyed the fruits of their labors as individual wage earners and consumers. Only 29 percent kept *any* of their annual earnings. The remaining 71 percent handed all of their wages over to their parents. That their earnings were treated exclusively as part of the family wage speaks volumes about the relationship between mill work and female empowerment in the early twentieth century South. For those who kept some of their annual earnings, the scores for SHARE are curiously distributed across a wide range of values, ranging

from a minimum of $13 (less than 1 percent of that worker's annual earnings) to a maximum of $294 (87 percent of her earnings), with the distribution trailing downward after the $200 mark.

[Table 1 about here]

The original question I intended to ask was simple enough: Which working girls got to keep more? Did individual characteristics determine how much money a worker retained for herself? If so, that would indicate an increasing trend toward individual autonomy among working girls. Or, were family considerations paramount in calculating whether working girls kept their wages? If so, it would be difficult to postulate that textile work fostered any sort of female emancipation from the parental control. The initial goal of the study, then, was to determine wage-keeping functions, using SHARE as the dependent variable in an OLS model. But since it turned out that so many women kept nothing at all, a new question seemed more pertinent: What factors distinguished the minority of "keepers" from the rest? Asking which of the "keepers" got to keep more now seemed a separate question, albeit a related one. The analytical problem raised by the distribution of the dependent variable SHARE is illustrated by the diagram in Figure 3. The first stage of the problem (who got to keep something?) is essentially categorical. The second stage (among keepers, who got to keep more?) requires an analysis of interval-scale data.

[Figure 3 about here]

Given the data and the issues at hand, which regression method should be employed? A logit or probit model could effectively distinguish between keepers and non-keepers, and this information may be all we need to know. If individual characteristics did not determine the keeping of wages, but family characteristics did, we could argue that a young woman's ability to command her wages for her own use was not a matter of personal empowerment but of family structure or other familial circumstances. In some basic sense, this would answer the question we started with, namely, whether the industrialization of Georgia enhanced the socioeconomic opportunities available to young single women. But a logit would also restrict our understanding of the dynamics involved. It would, in effect, nullify a good deal of useful data. What about those who kept something? Why did some keep so little, while others kept nearly everything they earned? Were the factors that determined who kept more the same factors that distinguished keepers from non-keepers? One way to make use of this data and to get at these questions would be to build a two-step regression model. The first step would be a logit to differentiate the keepers. The second step would be to run an OLS equation for "keepers" only.

Econometricians might say, however, that this two-step strategy unnecessarily complicates matters, because the distribution of SHARE presents an apparent censorship problem. A tobit

model, they might say, would be the solution to the peculiar distribution of y. Unlike the logit model, tobit would preserve all of the original data. Instead of asking two separate questions, one with full data and the other with a subset of that data, tobit would return us to a single query: Who kept more? At the same time, the question of who was more likely to keep anything at all would not be lost, since it would be inherent in the tobit results.

But is there really a censorship problem? The answer is not entirely clear. If we assume that zero is indeed the "actual" value for all non-keepers, no censorship problem exists. The status of a non-keeper, reflected in zero, is indeed the very thing we wish to measure. Given these assumptions we might argue that despite the skewed distribution, no hidden values are obscured from view, and OLS may be used. But what if we assume that zero does not necessarily reflect the same thing for all non-keepers? Zero, after all, indicates only that a working girl kept nothing; the assumption that she therefore has no capability of keeping some wages may well be false. It is not difficult to imagine situations in which zero is an inaccurate indication of a woman's potential for keeping wages. Suppose, for example, that a girl who normally retained some of her wages kept nothing in 1907 due to an unprecedented family crisis. Or suppose a woman with strong potential for keeping wages put all her earnings in the household bank because the family was hoping to buy a farm and leave the factory behind. If the woman personally preferred life on the farm, she might make an individual decision to pool her own earnings with those of her parents and siblings in order to facilitate the move to the country. Of course, without data to measure the impact of family crises and individual desires, no multivariate statistical approach can ferret out such nuances. But the very possibility of these scenarios is a warning that the dependent variable is indeed censored.

Table 2 compares the results of various regression models based on the above considerations. The dependent variable for the two OLS equations and the tobit analysis is SHARE, defined as the actual or estimated amount of the wage a working girl was able to retain for herself. For the logit model, y is the discrete categorical variable KEEPERS, defined as follows: working girls who kept some wages=1; working girls who kept none of their wages=0. The independent variables are the same for each model. The first four independent variables represent individual characteristics of each working girl; the last five represent characteristics of her family. Definitions of the independent variables are provided in the table. The numbers in parentheses are t-scores. In this analysis, as well as the case study that follows, I considered t-scores with an absolute value of 2 or more as an indication of statistical significance (at slightly better than the .05 level). The first OLS model includes all 274 cases.[17] When compared with the tobit, it clearly underestimates the regression coefficients--badly so for some variables

(although in one peculiar instance--the variable INDEARN--the OLS coefficient is actually higher than the tobit). It is also a matter of concern that the tobit ascribed statistical significance for OLDSIB while the OLS model did not. The variable AGE offers another a comparison. While the OLS equation underestimates the coefficient, it essentially rates AGE as statistically significant; the tobit model suggests, on the other hand, that the influence of AGE on keeping wages should probably be ignored. Taken as a whole, these results may be interpreted to mean that the OLS model is suffering from a censorship problem. A firm commitment to OLS in this case does not therefore seem worth the risk.

The tobit, then, seems preferable to the OLS model, and its results offer a compelling story. Women who kept more of their annual earnings did so not because of individual traits--experience, drive, education, job skills. Rather, they got to keep more because of the nature of their families. The number of siblings in the household over sixteen years of age (OLDSIB), had a positive and significant impact on keeping wages. A father's occupation mattered as well. If a cotton mill girl's father also worked in textiles (TEXDAD), she was more likely to keep some of her earnings. Finally, a working girl's ability to keep some of her earnings was positively and significantly related to her family's total annual earnings (FAMEARN).[18] A briefly stated conclusion is that keeping one's earnings stemmed from nothing we associate with increased autonomy. The industrial family economy, even when it allowed single females to keep some of their earnings, did little to dislodge working girls from parental control or traditional family norms. Keeping wages was, for individual working girls, linked primarily to the structure of their families and their place within them, of their families' aggregate earnings capabilities, and of their fathers' occupational choice. Few single women textile workers kept any of their wages for themselves, and those who kept something did so for reasons unrelated to an enhanced notion of individualism.

How different are the results of the two-step approach? The logit model in Table 2 tells a story similar to that of the tobit, except that FAMEARN is not statistically significant. No individual characteristics in the logit model were related to a working girl's ability to retain some of her wages. Family structure and father's occupation were again decisive factors for "keepers." The first stage of this two-step approach thus parallels but does not precisely match the results of the tobit.

The second step, however, raises questions not brought to light by the tobit analysis. The second OLS model in Table 2 includes only the eighty working women who kept some of their yearly earnings. Among those who kept some wages, what personal or familial characteristics determined who got to keep more? In this model, neither OLDSIB nor TEXDAD are statistically significant. While those variables separated the keepers from the non-keepers in both the tobit

and the logit, they had no significant effect on how much a "keeper" was able to retain. Instead, two other variables -- FAMEARN and INDEARN -- drove the bargaining power of the wage keepers. Notice the variable FAMEARN. It was statistically significant in the tobit but proved insignificant in the logit; then, in the second-stage of the two-step model, its again emerges as statistically significant. The implication is that the tobit result for FAMEARN is reflected in the second-stage of the logit/OLS method. If such an interpretation is correct, the two approaches seem to square with each other.

But not quite. What about the coefficient for the variable INDEARN (defined as a woman's gross annual earnings)? This is the only coefficient for any of the models that marks an individual characteristic as being significantly related to y. As the very high t-statistic indicates, this explanatory variable had a undeniable effect on how much of a woman's earnings she was able to command. As the coefficient reveals, for women who kept some of their earnings, the amount they got to keep increased 48 cents for every extra dollar they earned during the year. If a woman was in a position to keep some of her earnings, her personal earning ability made a positive difference in how much she kept. Since both INDEARN and FAMEARN are expressed in dollars, their coefficients in this OLS model may be compared. Since the amount a "keeper" was able to retain increased only 4 cents for every dollar the household members earned, it is clear that family earnings were not nearly as influential a determinate of y as a woman's individual earnings.

The two-step regression strategy thus suggests the following story. Whether a mill girl living at home was able to keep any of her wages did not depend on her work skills, her personal earnings power, or any other personal characteristic, such as experience on the job, age, or literacy. Instead, family structure and the father's occupation distinguished keepers from the rest. Yet, for those who kept something, individual abilities counted more than anything else in commanding how much was kept. When viewed narrowly in terms of keeping wages, the relationship between industrialization and female autonomy appears, for the great majority of working girls, to have been nonexistent. A select few, however, were able to have greater control over their economic lives than the rest, largely due to their own skill in making more money in the mill.

Which is the more appropriate model: the tobit or the two-step logit/OLS approach? There is no simple, technical answer to this question. Neither method violates the statistical guidelines for matching one's dependent variable with the proper type of regression equation. The ultimate evaluation of these approaches hinges less on technical issues than on the questions being asked. Both models suggest similar conclusions regarding industrial wage work and female

empowerment, and, had it not been for the INDEARN result in the stage-two OLS, the two models might be said to tell virtually identical tales. The statistical significance of INDEARN in the second-stage OLS model *does* set the two approaches apart, but it does not so much contradict the tobit model as it does extend our understanding of it. After all, the second-stage OLS model is dealing with a different set of data and asking a different question. The practical conclusion for historians is this: When faced with a peculiar dependent variable, the thoughtful experimentation with different types of regression models can strengthen and expand our understanding of the underlying dynamics involved in the observed relationships.

The second case study also deals with textile workers in Georgia, and it also involves a dependent variable with peculiar statistical qualities. In this instance, we focus specifically on the millhands of the Crown Cotton Mills of Dalton, a small textile town in the northwest part of the state. As part of a larger study of the Crown Mills community, I analyzed a four-month strike that occurred there in 1939. According to oral history interviews, the strike polarized the workers. For six weeks, union pickets kept the mill shut down; but then, suddenly, workers who opposed the strike broke through the picket lines and returned to work. For sixty working days--what I call the "critical period" of the strike--union loyalists stayed on the picket line even as antiunion millhands crossed the lines and put the mill back in partial operation. Ultimately, the strike ended in a stand off between the union workers and the company, with Crown Mill accepting the viability of the union and some of its demands and the union workers giving in to some wage cuts and work-load increases. My goal was to understand the division that tore Crown's millhands into warring factions during the strike of 1939. The workers' community there was closely knit; the workers knew each other well, and no outsiders were brought in as "scabs" to break the strike. The Crown Mill strike was an intense, internal conflict among working-class families, and I sought to understand that conflict.[19]

From a company payroll book I was able to determine how many days each worker remained on strike during the critical period. I therefore needed a statistical means of disentangling the various factors that went into the workers' strike decisions. Multiple regression offered an obvious solution. It would allow me to determine what sort of millhands were most likely to remain loyal to the strike. Determining this would help me resolve several important debates in the field of southern labor history and test some basic neoclassical economic assumptions about why workers go on strike.

Thinking about my regression strategy in advance, I intended to use as my dependent variable the total number of days each worker remained on strike during the critical period. That

number of days--the variable SUPPORT (0 through 60)--should have been a perfectly appropriate variable for an OLS model. But as Figure 4 illustrates, the values of SUPPORT were peculiarly polarized. During the critical period, workers did not trickle back to work individually. Rather, the millhands who crossed the picket lines did so as a group almost as soon as the lines broke. And millhands who supported the strike in the first days of the critical period proved to be diehard loyalists, staying out until the bitter end. Very few workers appeared to be neutral. Fence-sitters and opportunists, those waiting to see which way the wind would blow, seem to have been few.

The distribution of SUPPORT created potential problems for the OLS method. Although the data represents a standard interval-scale variable, the severely polarized distribution raised the question of whether a variable so distributed would function mathematically in an OLS equation as a dummy variable. If so, the OLS estimates might well be inaccurate, especially since statisticians generally agree that OLS is particularly sensitive to the problem of discrete dependent variables when individual-level data is used. It therefore seemed to me that following my initial strategy (an OLS model using SUPPORT as y) was inappropriate. Given the circumstances, the odds of statistically spurious coefficients appeared to be quite high.

One solution was to take the hint offered by the distribution of SUPPORT and shift to a logit or probit strategy. This required the creation of a truly dichotomous variable that distinguished strike supporters from strikebreakers. I therefore used the values of SUPPORT to create the variable DIEHARD, which took the value of 1 if a worker remained on strike 54 days or more during the critical period, and 0 otherwise. When making a dummy variable from an interval-scale variable there is no magic formula for marking the cut off point. Using 54 days as the dividing line for diehards was an arbitrary decision based on the distribution, which showed the diehard strikers as a clearly demarcated group at the high end of the scale. From 39 days into the critical period until 55 days, virtually no one crossed the picket line, so I could just as well have made the cut off point 40 days. On day 55, twelve millhands (the least devoted of the diehards) returned to work; that was the first day since day 11 of the strike that more than ten millhands crossed the picket lines. What about the small minority of workers who crossed the picket lines sometime between day 11 and day 55? My decision to place them with the strikebreaker group was also based on the distribution of SUPPORT, which showed that the diehards were a more cohesive group than the strikebreakers; hence, my decision to segregate the diehards from the rest in the regression model, and to combine the small group of "neutrals" with the more ardent strikebreakers.

Table 3 shows comparative models using logit, probit, and OLS methods. DIEHARD was the dependent variable in the logit and probit equations. The OLS model used SUPPORT as the dependent variable. These three models all utilize the same five independent variables. EARNINGS is a millhand's *normal* earnings capability prior to the strike. It is intended to test the neoclassical economic assumption that material considerations determined individual strike decisions. TENURE is a rough index of how long a millhand had worked for the company prior to the strike. It allows us to determine whether long-time workers behaved differently than newcomers during the strike. About one-third of the Crown Mill work force was female, and GENDER examines a basic characteristic that might have influenced strike behavior. Data on marriage and age were unfortunately not available. The variables for workers in MILL 1 and BOYLSTON are both dummy variables that indicate which of Crown's three plants the millhands worked in. Crown's three plants were the Mill No. 1 and Mill No. 2, which stood side by side in north Dalton, and the Boylston Mill, which was located one mile south of the others. As always, the key in using dummy variables to represent multiple categories is to include n-1 categories in the equation. Hence, only two of Crown's three plants are explicitly represented in the equation. The Mill No. 2 is implicitly represented in the Constant term, since all of the coefficients may be interpreted in relation to the Constant.

[Table 3 about here]

As anticipated, the figures show that logit and probit do indeed yield similar (though not equivalent) results. The regression coefficients in the OLS equation seem to jibe with the two nonlinear models, but the OLS results, it turns out, are suspect. In the logit analysis, GENDER and the millhands' place of employment (MILL 1 and BOYLSTON) prove to be significantly related to strike support. Men were more likely to be diehard strikers than women. Millhands in Mill No. 1 were more likely than those in the other mills to be diehards; those in Boylston were most likely to be strikebreakers. EARNINGS and TENURE prove insignificant in determining strike support. In the probit model, the estimated coefficients are consistently lower than the logit coefficients but not dramatically so. Moreover, all of the signs are the same and the t-scores are nearly indistinguishable. The "percent correctly predicted," the nonlinear analogue to R^2 statistic, is identical for both models.

The OLS results present an interesting contrast. Least-squares coefficients are not comparable to the nonlinear coefficients, since the dependent variables and the methods of calculation are different. But we *can* ask whether the nonlinear results and the OLS model tell essentially the same story. To some extent, they do. The t-scores pinpoint the same variables as significant. (One exception is that the CONSTANT term is significant in the OLS model, but this

result has no useful interpretive qualities; that is why quantifiers sometimes fail to report a t-score for the CONSTANT). So, even with the oddly distributed dependent variable, the OLS method seems generally robust. But the coefficient for EARNLOG is cause for suspicion, since the sign differs from that in the logit and probit. And even more problematic is the R^2 statistic, which, at .06, is suspiciously low.

As most readers will know, the R^2 statistic indicates how well the independent variables in an OLS equation actually explain variations in *y*. It is a number between 0 and 1 which indicates how well the model "fits" the data. The higher the number the better the fit. When using individual-level data, historians should always be prepared for low R^2 scores. But very low scores raise concerns about the viability of the equation. What is a "very low" R^2? There is no technical cut off point for making such a determination, but one very seldom sees regression models, even with individual-level data, with an R^2 of less than .15. Since the same independent variables give solid results and high levels of explanation in the logit and probit models, we may assume that these variables should explain *y* sufficiently well. Since they do not, the prudent assumption is that the peculiar distribution of the dependent variable does in fact damage the OLS results.

However robust the OLS estimates seem, then, the safest bet in this case is to reject the least squares approach and to use logit or probit method instead. This strategy places limits on our predictive ability. The OLS in Table 3 indicates that workers in Mill No. 1, on average, tended to remain on strike 9.5 days longer than workers in the other mills. Similar predictive conclusions cannot be made with coefficients estimated by logit and probit. But since the OLS model is suspect, we are better off narrowing our investigation to a more basic question--which side were the workers on during the strike--and using a nonlinear approach with a categorical dependent variable.

These two case studies, briefly explicated, can serve to underscore several interrelated points. Of the many considerations that go into any multivariate regression analysis, choosing the dependent variable is clearly the most fundamental. As these studies show, that choice is not always simple or self-evident. We must think hard in advance about the question we are asking, for the question itself determines our choice of *y*. But the data may force some reconsideration of our initial strategy and may force us to ask a different question. No social-scientific inquiry is as neat as the statistical textbooks suggest. Those texts simplify for the sake of lucidity; they stick to the basics in order to convey basic concepts. In the real world of research, for economists as well as historians, data sets are sometimes chaotic and more often than not they do not quite measure exactly what we want. As potters confront their lump of clay and begin to shape it, social

scientists confront their data and begin to work with it in conventional ways. But just as the clay itself sometimes governs the shape of a pot, the data we use often alters the nature of our investigations.

For this very reason, it is critically important that we acquaint ourselves with these four basic regression methods. Understanding when and why to use tobit, logit, and probit allows us to extend the range of data we can use, since any number of potential dependent variables that are unsuitable for OLS will be perfectly acceptable for one of the alternative methods. Choosing among these methods seems, superficially, an easy and predictable task. But when the nature or distribution of the dependent variable is peculiar, as it is in the investigations presented here, the choice is not always so obvious.

In certain cases, the method we choose may boil down to matters more philosophical than technical: What is being asked of the data and why? What sort of coefficients will be most useful to us? How much do we know about the data, about how it was collected, and how it tends to operate in a wide variety of regression models? Non-quantifiers might be surprised that this sounds more artistic than scientific, but experienced quantifiers will appreciate the point. The key point is that a broader menu of choices, when considered intelligently, can ultimately bolster our confidence in the regression models we present. When two or more methods appear to be statistically valid and of equal value, the most honest and useful approach is to be explicit in explaining the research design and to present as many different models as readers need to see.

The willingness and ability of historians to utilize regression techniques has now far surpassed the initial forays of the early 1970s. But if it is encouraging that more and more of us are using multiple regression, it is also imperative that we regress well. There are other issues regarding the relationship between dependent variables and regression strategies--such as the issue of "self-selectivity" in limited dependent variables--that historians need to discuss, and perhaps others will continue to advance those topics in the literature of historical methodology. Hopefully this essay will at least encourage others to use tobit, probit, and logit, and to appreciate the potential complications inherent in selecting and using dependent variables in multivariate regression analyses.

Table 1: Wages kept by Georgia's "Cotton Mill Girls" in 1907

Amount Kept (in dollars)	Working Girls in Each Category (%)
-0-	194 (70.8)
1 - 50	18 (6.6)
51 - 100	26 (9.5)
101 - 150	13 (4.7)
151 - 200	15 (5.5)
201 +	8 (2.9)
	N = 274 (100.0)

Table 2. Comparative Regression Models: Factors Related to Keeping Wages

	OLS [share]	Tobit [share]	Logit [keepers]	OLS: Keepers Only [share]
AGE	1.82	4.79	.06	1.13
	(1.94)	(1.77)	(1.67)	(.79)
INDEARN	.06	.03	-.00	.48
	(1.38)	(.28)	(1.76)	(6.25)
LITERACY	-13.33	-30.07	-.27	-7.62
	(1.64)	(1.22)	(.82)	(.54)
WEAVER	8.80	32.71	.43	2.29
	(.92)	(1.15)	(1.09)	(.14)
FAMEARN	.03	.07	.00	.04
	(2.77)	(2.27)	(1.30)	(3.14)
YOUNGSIB	-1.60	-4.27	-.04	-3.41
	(.81)	(.71)	(.52)	(1.07)
OLDSIB	5.89	23.22	.48	-9.91
	(1.55)	(2.14)	(2.93)	(1.80)
TEXDAD	28.08	85.23	1.44	-16.86
	(3.29)	(3.31)	(3.91)	(1.16)
DADHOME	-12.62	-44.20	- .76	-20.58
	(1.35)	(1.50)	(1.84)	(1.24)
Constant	-53.75	-263.25	-2.95	-58.23
	R^2 = .18		PCP = 77%	R^2 = .55
	N = 274	N = 274	N = 274	N = 80

(table continued)

(Table 2. cont.)

Dependent Variables [signified in brackets below type of method]:

SHARE = Estimated annual earnings a working girl got to keep, in dollars

KEEPERS = 1 if working girl kept any of her annual wages; 0 otherwise.

Variables Measuring Individual Characteristics

AGE = in years; only women under 40 years of age included (two cases in sample excluded by this criteria)

INDEARN = Total annual earnings of each working girl, in dollars.

LITERACY = 1 if read and writes; 0 otherwise.

WEAVER = 1 if works as weaver; 0 otherwise.

Variables Measuring Family Characteristics

FAMEARN = Total annual family income for girl's family, in dollars.

YOUNGSIB = Number of siblings living at home under 16 years of age.

OLDSIB = Number of siblings living at home who were 16 or older.

TEXDAD = 1 if girl's father was a textile worker; 0 otherwise.

DADHOME = 1 if girl's father lived at home; 0 otherwise.

Source of Data: U.S. Senate, Report on the Condition of Women and Child Wage-Earners in the United States, GPO, 1910, Vol 1, Cotton Textile Industry, pp. 934-1013.

Table 3. Comparative Regression Models of Strike Support

	Logit[1]	Probit[1]	OLS[2]
Earnings	.21 (.34)	.15 (.38)	- 4.39 (.59)
Tenure at Crown	- .11 (.57)	- .06 (.59)	- 2.14 (.96)
Gender (Male)	.41 (2.30)	.26 (2.33)	4.26 (1.98)
Worker in Mill No. 1	.80 (3.42)	.50 (3.51)	9.53 (3.61)
Worker in Boylston	- .72 (3.15)	- .45 (3.19)	- 11.58 (4.30)
Constant	- .55 (.42)	- .38 (.47)	43.45 (2.73)
	PCP=58%	PCP=58%	R^2= .06

N = 627
t - Scores in parentheses
[1] Dependent Variable: Diehard (on strike >53 days = 1; otherwise 0)
[2] Dependent Variable: Support (n days on strike)

Figure 1. Effect of Truncation on Regression Line

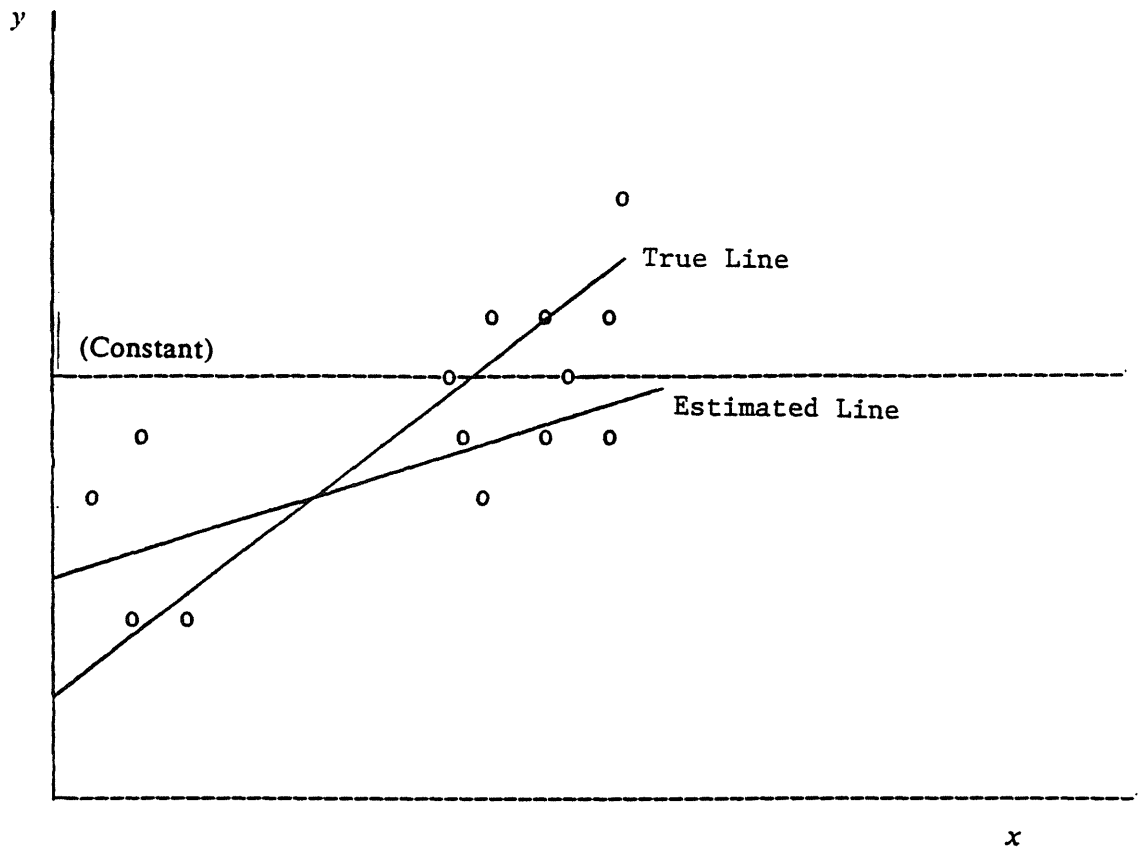
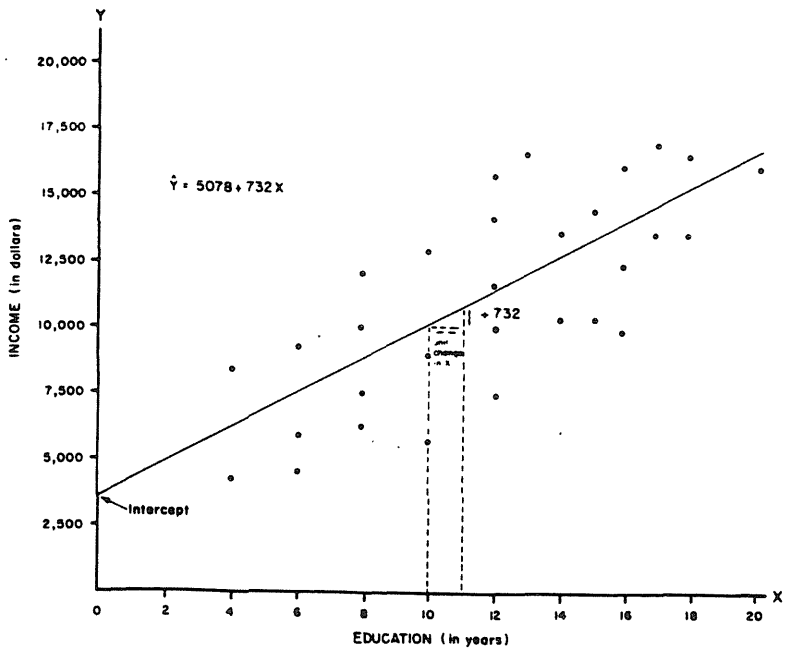


Figure adapted from Maddala, Limited-Dependent and Qualitative Variables in Economics, p. 167.

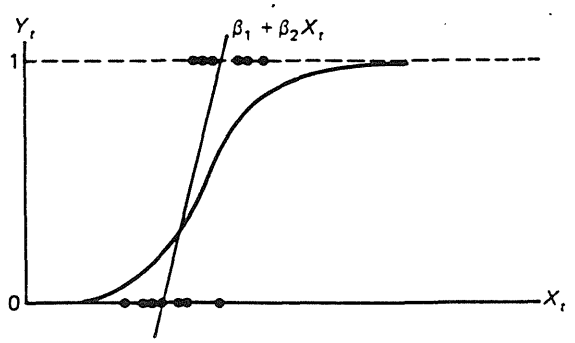
Figure 2. Bivariate OLS and Logit Models

a. A standard OLS Model



Note: Reprinted from Lewis-Beck, Applied Regression, p. 18.

b. A hypothetical logit model with S-curve and an OLS regression line.



Note: Reprinted from Hanushek and Jackson, Statistical Methods, p. 186.

Figure 3. Georgia's "Cotton Mill Girls" and Their Wages, 1907

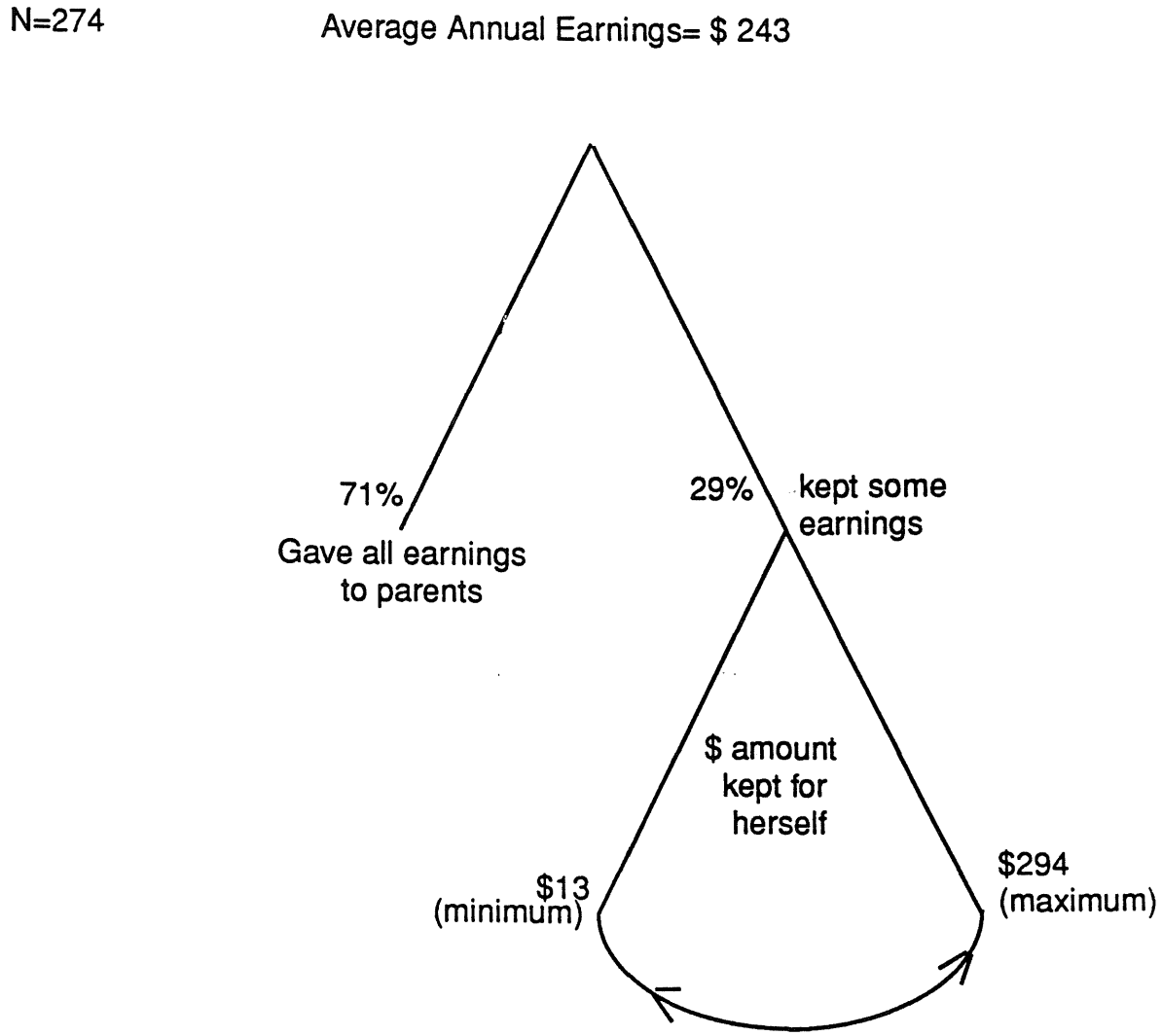
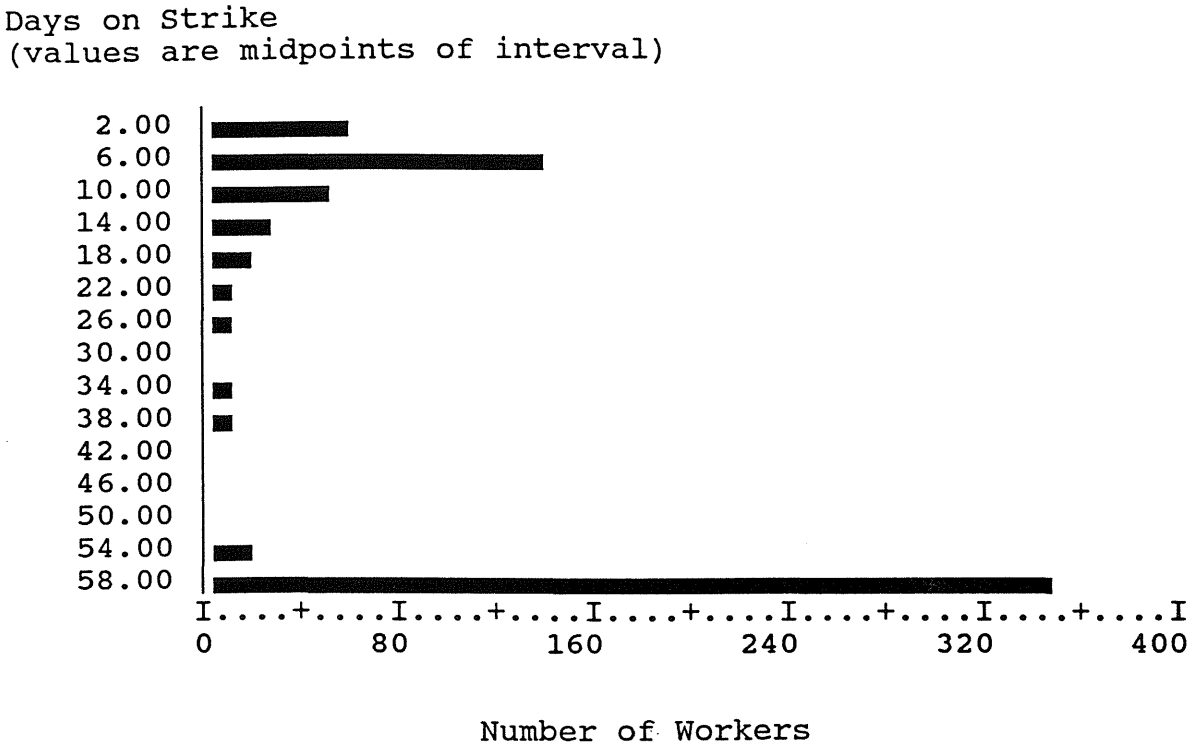


Figure 4. Strike Support Among Crown Mill Workers, 1939



ENDNOTES

1. Much of the initial use of, and debate over, multiple regression centered on the "new political history," in which some scholars used "ecological regression" strategies to discern patterns of mass voting behavior. See, E. Terrence Jones, "Ecological Inference and Electoral Analysis," *Journal of Interdisciplinary History* 2 (Winter 1972): 249-62; J. Morgan Kousser, "Ecological Regression and the Analysis of Past Politics," *Journal of Interdisciplinary History* 4 (Autumn 1973): 237-62; idem., *The Shaping of Southern Politics: Suffrage Restriction and the Establishment of the ONe-Party South, 1880-1910* (New Haven: Yale Univ. Press, 1974); idem., "The 'New Political History': A Methodological Critique," *Reviews in American History* 4 (March 1976): 1-14; Allan J. Lichtman, "Correlation, Regression, and the Ecological Fallacy: A Critique," *Journal of Interdisciplinary History* 4 (Winter 1974): 417-33.

2. E.g., Lee Benson, "The Mistransference Fallacy in Explanations of Human Behavior," *Historical Methods* 17 (1984): 118-31; and see the rebuttal by J. Morgan Kousser, "Must Historians Regress?: An Answer to Lee Benson," *Historical Methods* 19 (1986): 62-81.

3. Philip T. Hoffman, *Church and Community in the Diocese of Lyon, 1500-1789* (New Haven: Yale Univ. Press, 1984); Leonard K. Moore, *Citizen Klansman: The Ku Klux Klan in Indiana, 1921-1928* (Chapel Hill: Univ. of North Carolina Press, 1991); Daniel Scott Smith, "'All in Some Degree Related to Each Other': A Demographic and Comparative Resolution of the Anomoly of New England Kinship," *American Historical Review* 94 (1989): 44-79, esp. 66-67; Dale Baum, *The Civil War Party System: The Case of Massachusetts, 1848-1876* (Chapel Hill: Univ of North Carolina Press, 1984); William E. Gienapp, *The Origins of the Republican Party, 1852-1856* (New York: Oxford Univ. Press, 1987); John F. Reynolds, *Testing Democracy: Electoral Behavior and Progressive Reform in New Jersey, 1880-1920* (Chapel Hill: Univ. of North Carolina Press, 1988); Laird Boswell, "The French Rural Communist Electorate in the 1920s and 1930s," *Journal of Interdisciplinary History* (forthcoming, vol. 23). I use both OLS and logit in *Creating the Modern South: Millhands and Managers in Dalton, Georgia, 1884-1984* (Chapel Hill: Univ. of North Carolina Press, forthcoming, January 1993), chs. 5, 10.

4. When political historians use "ecological" regression models to explore voting behavior, they are utilizing OLS. The term ecological refers to the data and the assumptions made about the data, not the type of regression equation. Similarly, political historians' "transition tables," which measure party allegiance from one election to another, are also based on OLS equations. A

pioneering use of tobit in historical scholarship was Philip T. Hoffman, "Wills and Statistics: Tobit Analysis and the Counter Reformation in Lyon," *Journal of Interdisciplinary History* 14 (1984): 813-834.

5. Useful introductions to OLS regression include: Michael S. Lewis-Beck, *Applied Regression: An Introduction* (Beverly Hills: Sage Publications, 1980); Loren Haskins and K. Jeffrey, *Understanding Quantitative History* (Cambridge: MIT Press, 1990). Highly technical discussions include: E. A. Hanushek J. E. Jackson, *Statistical Methods for Social Scientists* (New York: Academic Press, 1977), chs. 2-4; and G. S. Maddala, *Econometrics* (New York: McGraw-Hill Book Company, 1977), chs. 6-11.

6. In my view, the statistical package that runs all four types of regressions most conveniently is SST, developed at Caltech by Jeffrey A. Dubin and R. Douglas Rivers (software and manuals may be obtained via professor Dubin, Division of Humanities and Social Sciences, 228-77, Caltech, Pasadena, CA 91125). The mainframe program SPSS-X offers an alternative, log-linear form of logit.

7. Multicollinearity means that one explanatory variable in the regression model is highly correlated with another. In such a case, the estimated coefficients will likely be unreliable. We will not be able to compare, with any confidence, the relative effects of the explanatory variables, and our significance scores may be deceptively low (i.e., whereas an actual relationship exists between independent variable x and dependent variable y, the t-score for x may be driven so low by the multicollinearity problem that one will conclude that no statistically significant relationship exists). How to determine whether multicollinearity exists and what to do about it is discussed elsewhere, very usefully in Lewis-Beck, *Applied Regression*, pp. 58-63.

8. Sometimes the dependent variable needs to be altered in order for the equation to work more effectively. For example, economists often replace the actual values of y with a logarithmic transformation of y. This is done to improve the distribution of y for purposes of calculation. "Logging" the dependent variable does not alter the basic form or function of an OLS regression equation.

9. G. S. Maddala, *Limited-Dependent and Qualitative Variables in Econometrics* (Cambridge: Cambridge Univ. Press, 1983), chs. 1, 6.

10. Another useful example of censorship (suggested by Hoffman, Wills and Statistics, p. 832) is the demographic issue concerning the age at which people marry. If, by law, people must reach a

certain age before they can be wed, then the dependent variable "age at marriage" will be artificially censored at the lower end of the scale, in accordance to the prevailing law.

11. Maddala, *Limited-Dependent*, pp. 1-4, 165-70.

12. Maddala. *ibid.*, p. 151, notes that the name of the procedure was derived from James Tobin, "who first discussed this problem [of truncation] in the regression context" and did so in the context of probit analysis. "Tobin's probit," as one researcher called it, became tobit. Hoffman. *Church and Community*, pp. 171-84, offers a lucid introduction to tobit. The initial methodological piece is James Tobin, "Estimation of Relationships for Limited Dependent Variables," *Econometrica* 26 (1958): 24-36.

13. Maddala, *Limited-Dependent*, ch. 2; Hanushek and Jackson, *Statistical Methods*, ch. 7, esp. 184-86.

14. D. Roderick Kiewiet, *Macroeconomics and Micropolitics: The Electoral Effects of Economic Issues* (Chicago: Univ. of Chicago Press, 1983). pp. 139-41; on the conversion of tobit estimates, see, Hoffman, *Church and Community*, pp. 176-77.

15. On the debate over the meaning of women's work and women's emancipation, see, Edward Shorter, "Female Emancipation, Birth Control, and Fertility in European History," *American Historical Review* 78 (June 1973): 605-640, esp. 612-17; Joan W. Scott and Louise A. Tilly, "Women's Work and the Family in Nineteenth Century Europe," *Comparative Studies in Society and History*. 17 (1975): 36-64; idem., *Women, Work, and Family* (New York: Holt, Rinehart and Winston, 1978); Leslie Tentler, *Wage-Earning Women: Industrial Work and Family Life in the United States, 1900-1930* (New York: Cambridge Univ. Press, 1979). A quantitative analysis of the issue is Gary Cross and Peter R. Shergold, "The Family Economy and the Market: Wages and Residence of Pennsylvania Women in the 1890s," *Journal of Family History* 11 (1986): 245-65.

16. U.S. Senate. *Report on Condition of Woman and Child Wage-Earners in the United States* (Doc. no. 645, 61th Congress, GPO, 1910), vol. 1, Cotton Textile Industry, Table XXX (data arranged by state), pp. 934-1013; on the collection and reliability of the data (the government agents state that the data on annual earnings, though carefully gathered, is "necessarily only approximate"), see pp. 932-35.

17. The *n* in these regression equations is 274 instead of the original 276 because I excluded the two single women in the sample who were over forty years of age (what mill people would have called "spinsters") on the grounds that their behavior might well have been shaped by different imperatives than the younger women. Technically, this exclusion amounted to a truncation of the

data, but because the *n* excluded was so small the effect was trivial and, in practice, did not alter the results in any way.

18. The number of older siblings naturally had some bearing on family earnings; the simple correlation between the two was .60; TEXDAD, surprisingly, had virtually no relationship to FAMEARN.

19. Crown Cotton Mills Payroll Books, 1939, Crown Gardens and Archives, Dalton, Georgia. For more on the data used in the regression here, see, Flamming, *Creating the Modern South*, Appendixes.