#430-77

INSTITUTE FOR RESEARCH ON POVERTY DISCUSSION PAPERS

ESTIMATION WHEN THE SAMPLING RATIO IS A LINEAR FUNCTION OF THE DEPENDENT VARIABLE

John Bishop

UNIVERSITY OF WISCONSIN ~ MADISON

Estimation When the Sampling Ratio
Is a Linear Function of the Dependent Variable

John Bishop

August 1977

The research reported here was supported by funds granted to the Institute for Research on Poverty at the University of Wisconsin by the Office of Economic Opportunity pursuant to the provisions of the Economic Opportunity Act of 1964. The conclusions expressed herein are those of the author.

ABSTRACT

Many of the data sets used by economists and sociologists to estimate relationships between social origins of a youth and his later success in life suffer from a serious nonresponse bias, where not responding is negatively associated with success. An expression for the bias in ordinary least squares (OLS) estimated coefficients is calculated for the case where the probability of being in the sample is a linear function of the dependent variable. Adopting the conventional assumption that the true relationship has a homoskedastic error structure, we find that the ratio of the true to the estimated coefficient of a bivariate and trivariate regression is a simple positive function of the R^2 of the true relationship and a negative function of the absolute size of the proportionate change in sampling probability for a standard deviation change in the dependent variable. When independent variables are symmetric, the bias is independent of whether the sampling proportion is a positive or negative function of the dependent variable.

Estimation When the Sampling Ratio Is a Linear Function of the Dependent Variable

It is not uncommon for economists and sociologists to use data bases where the probability that a random individual will be in the sample depends upon his income, occupation, or education. Often these data bases are used to estimate models predicting these same success indicators. The application of ordinary least squares (OLS) to such data, however, yields inconsistent estimators of models predicting income, occupation, or education.

The biased nature of OLS estimators when the sample selection is based on the dependent variable, often called truncation bias in the literature, has been pointed out frequently (Bishop 1974; Cain 1975; Crawford 1975; Hausman and Wise 1977; Manski and Lerman 1976; Taubman and Wales 1974, ch. 4, app. F, L). Sometimes the sampling process results in an absolute truncation (i.e., absolutely no one with initial year incomes above 1.5 times the poverty line, as in the Rural Income Maintenance Experiment). Estimation techniques for this situation have been developed by Crawford (1975) and Hausman and Wise (1977).

This paper tackles the situation where all observations in the population have some probability of being in the sample and the probability is a linear function of the dependent variable. I calculate and apply a formulae that relates the bias to the strength of success selectivity and the R^2 of the true relationship.

Data bases where sampling ratio depends upon income are of two types: follow-up surveys with substantial nonresponse rates, and interview surveys that oversample people in low or high income neighborhoods. Follow-up surveys may fail to obtain information from many of the people in its defined sample for a variety of reasons: death, inability to find a current address, or refusal by the respondent to fill out the questionnaire. Refusals are the primary cause of success bias.

One heavily used data set with substantial refusal problem is Project Talent. The combined 1 and 5 year follow-ups of the male 11th graders had a response rate of 52% to the series of mail questionnaires. A special intensive follow-up of a 5% sample of mail questionnaire nonrespondents which obtained a 90% response rate allows us to establish the extent to which success affects the probability of responding to a mail questionnaire. Stratifying by the social status of each student's parents, college attenders were 1.5 to 1.6 times as likely to respond to at least one of the two follow-ups (Bishop 1974). Given college attendance status, the student's family background had no systematic impact on his response rate.

Another very important data set that potentially has a success bias is the National Bureau of Economic Research (NBER)-Thorndike sample. Thorndike took a random sample of 17,000 from a population of Army Air Corps volunteers for pilot, navigator, and bombardier training programs who passed a preliminary screening test. By 1955, 1500 had died and of the living, 2000 military and 9700 civilians responded to a mail questionnaire. The response rate was therefore about 75%. This is a high response rate and is attributed by Taubman and Wales (1974) to the accurate current addresses generally available from the Veterans Administration and the use of Retail Credit Bureau to find some of the nonrespondents.

The 1969 data is a survey of the 1955 respondents. Of those for whom current addresses were obtained and who had not died, 70% responded. Taubman and Wales found that while the 1955 income of 1969 nonrespondents was lower than for the respondents, it was not lower when ability and schooling were controlled. From this they argued that any selection process that existed was based on the independent and not the dependent variables. It has been shown that when the true model has homogeneous coefficients, differential sampling ratios that depend on included right hand side variables do not bias the estimates of structural parameters (Porter 1973; Taubman and Wales 1974).

However, their test applies only to the response rate conditional upon having responded in 1955. There may still be success bias in the 1955 response rate. Their test also depends upon the assumption that success persists over time and that income is as good a measure of success at age 29 as at age 44. If the conditional probability of responding in 1969, given that one responded in 1955, is a function of the change in one's relative income over the period, the test used by Taubman and Wales will miss the success bias. An alternative way to test for success bias in the 1969 data would be to compare those who responded as soon as they received a questionnaire to those who required reminders. But even this requires some strong assumptions. Because of the lack of an intensive follow-up by retail credit or phone, we can never be sure there is no success bias in the NBER-Thorndike data. However, it may be possible to put limits on the effects a success bias could have.

Another type of data set in which this problem arises is when black neighborhoods have been oversampled, as in the 1966-67 Survey of Economic

Opportunity (SEØ); when low income neighborhoods have been oversampled, as in the Census Employment Surveys; or when low family incomes relative to the poverty line are oversampled, as in the Michigan Panel Study of Income Dynamics. These data sets have been used to estimate models predicting success variables like hours worked, weeks worked, and earnings. A widely publicized finding using these surveys has been that rates of return to schooling are lower in low income neighborhoods than for samples of people drawn from the metropolitan area as a whole or the nation (Harrison 1972). Since living in a poverty neighborhood is a consequence of earnings, restricting one's sample to these neighborhoods or oversampling in them results in a simultaneous equations bias when estimating the structural parameters of models that predict earnings and other success variables.

In the next section of this paper, I calculate the bias to be expected in OLS estimates of structural models of earnings, work effort, or status attainment when the probability of being in the sample is a linear function of the dependent variable. If we adopt the conventional assumption that the true relationship has a homoskedastic error structure, we find that the ratio of the true to the estimated coefficient is a simple positive function of the R^2 of the true relationship and a negative function of the absolute size of the proportionate change in sampling probability for a standard deviation change in the dependent variable. When the right hand side variables are symmetric (the third moment = 0), the bias is independent of whether the sampling proportion is a positive or negative function of the dependent variable. To demonstrate the importance and relevance of these findings, the final section of this paper compares the schooling coefficients estimated in different subsamples of the SEO in models predicting yearly earnings.

1. Statistical Model

Porter (1973) and others have shown that if sampling ratios are independent of the disturbances of the model to be estimated and the coefficients of that model are homogeneous over the population, OLS estimatoors of structural parameters are unbiased. In other words, sampling ratios that are functions of included independent variables (correlated with y only because of the joint dependence of x and y) do not produce a selection bias in OLS estimators. The problem dealt with in this paper is sampling ratios that are linear functions of the dependent variable. Sampling proportions correlate with independent variables solely as a result of their joint association with y.

Analytical solutions are not difficult to obtain for models with only one independent variable. Let the true model be

1) $y_i = \beta x_i + u_i$

2) $p_i = (1 + \gamma y_i + v_i)\, n_s/n.$

Then

3) $$E_s(y) = \frac{\sum_1^n p_i y_i}{\sum_1^n p_i} = \frac{\sum_1^n (1 + \gamma y_i + v_i) y_i}{\sum_1^n (1 + \gamma y_i + v_i)} = \gamma V(y)$$

where i indexes each observation in the population (i = 1 . . . n)

y_i and x_i are defined as deviations from their population mean

u_i is homoskedastic and independent of x_i and v_i

p_i = probability the "i"th observation will be selected

n_s/n = the average sampling ratio = the number of observations selected for the sample (n_s) divided by the total number in the population (n)

v_i is independent of x_i and consequently independent of y_i

γ = the increased probability of being sampled per unit of y divided by the average sampling proportion

E is the expectation operator

s subscript indicates the mean, variance, or covariance indicated is for the nonrandom sample.

We note that all summations are over the entire population, i = 1 . . . n, and drop the limits from our notation. The sample mean of x is

$$4)\quad E_s(x) = \frac{\Sigma\, p_i x_i}{\Sigma\, p_i} = \frac{\Sigma\,(1 + \gamma y_i + v_i)\, x_i}{\Sigma\,(1 + \gamma y_i + v_i)} = \gamma\, \mathrm{Cov}(xy) = \gamma\beta V(x).$$

Noting that $\Sigma x = \Sigma y = 0$, the sample variances and covariances have the following expectations:[1]

$$E(V_s(x)) = \frac{\Sigma[1 + \gamma y_i + v_i][x_i - \gamma\beta V(x)]^2}{\Sigma(1 + \gamma y_i + v_i)}$$

$$= \frac{\Sigma[x_i - \gamma\beta V(x)]^2}{n} + \frac{\gamma}{n}\,\Sigma y[x - \gamma\beta V(x)]^2$$

$$= V(x) + \gamma^2\beta^2 V(x)^2 + \frac{\gamma\Sigma y_i x_i^2}{n} - 2\,\gamma^2\beta V(x)\mathrm{Cov}(xy)$$

$$5)\quad E(V_s(x)) = V(x)\left[1 + \frac{\gamma\beta\Sigma x^3}{nV(x)} - \gamma^2\beta^2 V(x)\right]$$

$$E(\mathrm{Cov}_s(xy)) = \frac{1}{n}\,\Sigma\,[x_i - \gamma\beta V(x)][y_i - \gamma V(y)]$$

$$+ \gamma\,\frac{1}{n}\,\Sigma y[x_i - \gamma\beta V(x)][y_i - \gamma V(y)]$$

$$= \mathrm{Cov}(xy) + \gamma^2\beta V(x)V(y) + \frac{\gamma}{n}\,\Sigma x_i y_i^2 - \gamma^2 V(y)\mathrm{Cov}(xy)$$

$$- \gamma^2\beta V(y)V(x)$$

$$6)\quad E(\mathrm{Cov}_s(xy)) = \mathrm{Cov}(xy)\left[1 + \frac{\gamma\beta^2\Sigma x^3}{n\ \mathrm{Cov}(xy)} - \gamma^2 V_{(y)}\right].$$

The probability limit of the sample estimate of β is

$$7)\quad b_s = \frac{\mathrm{Cov}(xy)[1 - \gamma^2 V(y) + \gamma\beta^2\Sigma x^3/n\mathrm{Cov}(xy)}{V(x)[1 - \gamma^2\beta V(x) + \gamma\beta\Sigma x^3/nV(x)}$$

$$8)\quad \frac{b_s}{\beta} = \frac{1 + D - \gamma^2\ V(y)}{1 + D - \gamma^2\ V(y)R^2},$$

where $D = \gamma\beta\Sigma x^3/nV(x) = \gamma\beta$ times the ratio of the third and second moments of x

R^2 = the proportion of the variance explained by the true relationship.

Since $R^2 \leq 1$, b_s/β is necessarily less than or equal to 1. Selection on the dependent variable attenuates the parameter estimates. The amount of attentuation depends upon three factors: the direction and degree of skewness of x(D), the strength of the relationship between y and the probability of selection (γ), and the R^2 of the underlying relationship.

The D term in (8) depends upon the interaction of the sample selection process with the skewness of x. Since skewness is defined as $a_3 = \Sigma x^3/n\sigma_x^3$ = the third moment of a variable over the cube of its standard deviation, we may rewrite $D = a_3 \cdot \gamma\beta\sigma_x = a_3 \cdot \gamma\sigma_y r_{xy}$. The expression , $\gamma\sigma_y r_{xy}$ times 100, can be interpreted as the percentage change in the probability of an observation's selection into the sample that is associated with a standard deviation change in x. It is positive when γ and r_{xy} have the same sign, as in earnings functions estimated on Project Talent or NBER-Thorndike data sets. Thus, if the distribution of x in the population has positive skew, D is positive, which reduces bias. In SEO and Census Employment Survey data sets where families in black or low income neighborhoods are oversampled, $\gamma\sigma_y r_{xy}$ is negative because here γ and r_{xy} have opposite signs. In these surveys a positive skew to x causes D to be negative, thus increasing the bias.

The distribution of years of schooling--the x variable upon which we are focussing in this paper--can be skewed in either direction, depending on the year and population studied. People educated in the early twentieth century have positively skewed educational attainment distributions. The most recent cohorts have negatively skewed distributions. Men between the ages of 30 and 35 in 1974 have an $a_3 = -.53$. Distributions for adults of all ages are very close to being symmetric. When compared to the skewness of a zero-one variable with a mean of .1, whose $a_3 = 2.67$, skewness for all adults is quite small: .04 for white males and -.14 for black males in the 1967 CPS. Since the term measuring the impact of a standard deviation change in x on the probability of selection, $\gamma\sigma_y r_{xy}$, must have an absolute value of substantially less than one, schoolings skewness does not have an important effect upon the magnitude of the selection bias in first order statistics of relationships between schooling and income. From this point on we will, therefore, neglect the impact of skewness and assume that all independent variables are symmetric ($a_3 = 0$). When all variables are assumed symmetric, it is possible to derive a simple formula for the selection bias in the coefficients of regressions with two independent variables. (The mathematical derivation is carried out in the Appendix.) The formula that results is the same as the formula for first order regression coefficients when x is symmetric:

$$9)\quad \frac{b_s}{\beta} = \frac{1 - \gamma^2 V(y)}{1 - \gamma^2 V(y) R^2},$$

where R^2 is the coefficient of determination in the multi or bivariate regression in the full population.[2] The sign of γ indicates whether the sampling ratio is positively or negatively associated with the dependent variable. It is squared in the final terms of both the numerator and denominator. Consequently, the size of the bias is not affected by whether more income raises or lowers the probability of selection. The probability limit of the ratio of estimated to true parameters when the independent variables are symmetric is presented in Table 1 for alternative γ's and R^2's.

If the $R^2 = 1.0$, there is no bias, for selecting the sample on the dependent variable is equivalent to selecting on the independent variables. As the R^2 declines, the bias increases in size for a $|\gamma\sigma_y|$ of .4, an R^2 of .6 implies a bias ratio of .929. An R^2 of .3 implies a bias ratio of .882 or a 12% attenuation of regression coefficients. An R^2 of .1 implies a bias ratio of .853 or a 15% attenuation of the coefficients. In the limit as R^2 approaches zero, the bias ratio approaches its maximum of $b_s/\beta = 1 - \gamma^2 V(y)$. Thus, when the bias in first order coefficients is compared across alternative right hand side variables, the proportionate attenuation is larger in variables that have a weak relationship with y. Since in a trivariate relationship bias depends upon the multiple correlation coefficient, the coefficients of both independent variables attenuate by an identical proportionate amount.

Table 1

Values of b_s/β as a Function of R^2 of the True Relationship and the Strength of Selection on y

	R^2 = 1.0	.8	.6	.5	.4	.3	.2	.10	0
$\|\gamma\sigma_y\|$ = .707	1	.833	.714	.667	.625	.588	.555	.526	.5
$\|\gamma\sigma_y\|$ = .5	1	.938	.882	.853	.833	.811	.789	.769	.75
$\|\gamma\sigma_y\|$ = .4	1	.965	.929	.913	.897	.882	.868	.853	.84
$\|\gamma\sigma_y\|$ = .2	1	.992	.984	.979	.975	.971	.967	.964	.96

Note: All independent variables are symmetric.

$\gamma\sigma_y$ is the proportionate increase in the sampling probability per standard deviation of the dependent variable.

The expression, $\gamma\sigma_y$, is the change in the probability of inclusion in the sample associated with a standard deviation change in y divided by the average probability of inclusion. The smaller $|\gamma\sigma_y|$ the smaller the bias. Since γ must approach zero as the proportion of a population that is sampled approaches one, selection bias must decline as a survey's response rate approaches 100%. For a given R^2, the attenuation of regression coefficients rises roughly in proportion to the square of $\gamma\sigma_y$. At an R^2 of .30, a $\gamma\sigma_y$ of .2 causes a 3% attenuation, a $\gamma\sigma_y$ of .4 causes an attentuation of 12%, and a $\gamma\sigma_y$ of .707 yields an attenuation of 41%.

Biases of even larger magnitudes are possible if selection probabilities have a nonlinear relation ($\ln \frac{P}{1-P} = \gamma y$, for instance) with the dependent variable. As long as the sampling ratio is defined as a linear function of y, it is not possible for our model to handle truly powerful selection biases. The derivations would be internally inconsistent if predicted sampling ratios fell outside the zero-one interval. They will not fall outside this interval if γ is sufficiently small and the y distribution sufficiently compact. A rectangular distribution for y would require a $|\gamma\sigma_y| < .81$, if $n_s|n < .5$, and a $|\gamma\sigma_y| < 2(.81)(1-n_s)/n$ for $n_s/n > .5$. All other single modal distributions of y will require that γ be smaller than these limits.

2. Application to Earnings Functions in the Survey of Economic Opportunity[9]

Our statistical model predicts that when the sampling ratio is dependent on income, the schooling coefficients in an earnings function will be lower than the true population coefficient. Table 2 tabulates estimated relative sampling ratios by earnings for alternative subsamples of the SEO. Not surprisingly, the probability of living in a low income neighborhood is negatively associated with the level of one's earnings. For whites, the probability of living in a predominantly black area is also negatively associated with earnings. For blacks, however, there was no visible relationship. Therefore, we do not expect blacks in the special sample of predominantly black neighborhoods to have lower schooling coefficients than a national sample of blacks. We do expect, however, that whites living in these neighborhoods will have a smaller schooling coefficient than a national sample of whites. Also, rates of return to schooling estimated for both blacks and whites living in low income neighborhoods in urban areas are expected to be smaller than the rates of return for all urban residents. An examination of Table 3 indicates, as expected, that schooling coefficients of whites in predominantly black and low income areas are substantially smaller than those in the national sample. For whites the unbiased coefficient of .0879 falls to .0701 in black areas and to .0643 when the sample is limited to low income neighborhoods. The schooling coefficients for blacks are smaller only for the low income areas. Furthermore, the drop in the schooling coefficients is larger for models with low R^2 (those without measures of work effort on the right hand side).

For blacks in low income areas the linear specification of the sampling mechanism predicts the coefficient changes well. For whites,

Table 2

Estimated Sampling Ratio Conditional Upon Income Relative to the Average Sampling Ratio

Earnings	0-2	2-3	3-4	4-5	5-6	6-7	7-8	8-10	10-14	14-20
Whites in predominantly black areas	2.24	2.24	2.47	1.53	1.41	.83	.90	.64	.48	*
Blacks in predominantly black areas	1.00	.91	1.02	.97	.83	1.09	.91	.88	*	*
Whites in low income areas	2.31	2.45	3.00	2.17	1.37	.87	.70	.56	.21	*
Blacks in low income areas[1]	1.07	1.08	1.27	1.12	1.12	.88	.63	.52	*	*

*means n of the Current Population Survey base is below 10.

[1] Since low income areas were defined only for Standard Metropolitan Statistical Areas the comparison base is all blacks living in SMSA's.

Table 3

Schooling Coefficients in Different Samples

	Low Income Area Coef.	Predom. Black Area Coef.	CPS			
			Coef.	R^2	$\gamma\sigma$* Predom. Black	$\gamma\sigma$* Low Income
Yearly Earnings						
Whites						
0-20 yrs schooling	.0643	.0701	.0879	.23	-.95	-1.06
0-15 yrs schooling	.0500	.0656	.0889	.17	-1.00	-1.11
Blacks						
0-20 yrs schooling	.0525	.0628	.0610	.08	.45	-.20
0-15 yrs schooling	.0447	.0530	.0621	.07	.45	-.20
Hourly Earnings, 0-20 yrs Schooling						
Whites	.0588	.0556	.0743	.40	-.95	-1.06
Blacks	.0410	.0504	.0462	.54	.45	-.20

Note: The dependent variable is the log of yearly earnings. Samples were limited to nonfarm males not in school with at least six years of experience. The schooling coefficients are from regressions with experience, experience squared, SMSA residence, and SMSA size as controls. The hourly earnings coefficients have additional controls: log of weeks worked last year, part time last year, and last week. The Black CPS sample was limited to SMSA residents.

*Estimates of γ were obtained from unweighted regressions of the ratio of the observed conditional sampling ratio to the average sampling ratio on the log of yearly earnings. The CPS provides the estimate of the population distribution of earnings. Weighted regressions yield more negative estimates of γ.

the impact of income on the sampling ratio is so powerful that the estimates of γ produced are too high. Some high earnings individuals will have negative predicted sampling ratios, in which case the analysis becomes internally inconsistent. If predicted coefficients are calculated, nevertheless, we overpredict the reduction in the schooling coefficients.

The problem is that for whites the sampling ratio-earnings relationship for predominantly black or low income neighborhoods is nonlinear. It looks like a logistic specification would serve better than a linear specification. Simple analytic results are not obtainable, however, when the sampling ratio is a nonlinear function of the dependent variable.

Where does that leave the researcher? If data availability forces one to use a data set in which sampling ratios are nonlinear functions of the dependent variable, how can consistent estimators be obtained? The solution that suggests itself is a two stage process. First, estimate a model of the sampling process. If sampling ratios depend directly on some of the independent variables as well as the dependent variable, these variables should be included in the model along with y. The main requirement of this model is that the error in predicting the sampling ratio be independent of the disturbances of the structural model. In Census Employment Surveys this could be done by comparing the low income area's population to that of the SMSA as a whole. In follow-up surveys a data set with an intensive follow-up of a sample of nonrespondents is required.

The second step is to estimate the structual model, using the inverse of these predicted sampling ratios as weights. Manski and Lerman

(1976) have shown that when probabilities of inclusion in the sample are a function of a categorical dependent variable, weighting each observation by the inverse of its sampling ratio yields unbiased and efficient estimators of the coefficients of a logistic model. Where sampling ratios are known (as for follow-up surveys with intensive follow-ups of a small sample) weighted least squares using these ratios from the sampling frame is another alternative. It is safe from misspecification of the sampling model but it becomes highly sensitive to the observations in the nonrespondent sample, since just a few observations carry a major share of the variance to be explained. Both approaches reduce bias only at the cost of increasing heteroskedasticity. The advantage of using predicted sampling ratios rather than sampling frame ratios is that the heteroskedasticity created by weighting will be less serious. Heteroskedasticity, however, does not bias coefficients, it only lowers the precision with which they are estimated.

This paper presents a suggested route for exploration. I leave the rigorous development of the properties of such estimators to a later time, and to others.

APPENDIX

Linear Selection Bias in the Trivariate Regression Model

Linear Selection Bias in the Trivariate Regression Model

In the true model,

1) $y_i = \alpha^* z_i + \beta^* x_i + \varepsilon_i$

2) $z_i = \phi x_i + v_i$

3) $x_i = \frac{Cov(xz)}{V(z)} z_i + v_i^*$

4) $p_i = (1 - \gamma y_i + u_i)\, n_s/n,$

where α^* and β^* are partial regression coefficients

ϕ, α, and β are first order regression coefficients

ε is independent of x, z, u, v_i and v_i^*

u is independent of x, z, y, v_i and v_i^*

γ = income selectivity parameter,

Note that

$E_s(y) = \gamma V(y)$ $E_s(x) = \gamma Cov(xy)$ $E_s(z) = \gamma Cov(zy)$

5) $E(V_s(x)) = V(x)[1 + D_x - \gamma^2 V(y) R^2_{yx}]$

6) $E(V_s(z)) = V(z)[1 + D_z - \gamma^2 V(y) R^2_{zy}]$

7) $E(Cov_s(xy)) = Cov(xy)[1 + D_x - \gamma^2 V(y)]$

8) $E[Cov_s(zy)] = Cov(zy)[1 + D_z - \gamma^2 V(y)],$

where $D_x = \gamma\beta\Sigma x^3/nV(x)$ if x is symmetric $D_x = 0$

$D_z = \gamma\alpha\Sigma z^3/nV(z)$ if z is symmetric $D_z = 0.$

Obtaining an expected value for the sample covariance of x and z is going to require that $\frac{\gamma\Sigma(xyz)}{n, Cov(xz)}$ be evaluated:

$$9)\quad \frac{\gamma}{n}\frac{\Sigma(xyz)}{Cov(xz)} = \gamma\Sigma xz(\beta^{*}x + \alpha^{*}z + \varepsilon)/n\ Cov(xz)$$

$$= [\beta^{*}\gamma\Sigma x^{2}z + \alpha^{*}\gamma\Sigma xz^{2} + \gamma\Sigma xz\varepsilon]/n\ Cov(xz)$$

$$10)\quad \frac{1}{n\ Cov(xz)}\ [\beta^{*}\gamma\ \frac{Cov(xz)}{V(x)}\ \Sigma x^{3}1 + \alpha^{*}\gamma\ \frac{Cov(xz)}{V(z)}\Sigma z^{3} + \beta^{*}\gamma\Sigma vx^{2} + \alpha^{*}\gamma\Sigma z^{2}v^{*} + \gamma\Sigma xz\varepsilon]\ .$$

The independence of x and v, z and v^{*}, and ε and x and z results in the last three terms being zero:

$$11)\quad \frac{\gamma}{n}\frac{\Sigma xyz}{cov(xz)} = \frac{\beta^{*}\gamma}{n}\frac{\Sigma x^{3}}{V(x)} + \frac{\alpha^{*}\gamma}{n}\frac{\Sigma z^{3}}{V(z)} = \frac{\beta^{*}D}{\beta^{x}} + \frac{\alpha^{*}D}{\alpha^{z}} = D_{x}^{*} + D_{z}^{*}$$

$$12)\quad E(Cov_{s}(xz)) = \frac{1}{n}\ \Sigma(x - \gamma Cov(xy))(z - \gamma Cov(zy))$$

$$+ \frac{\gamma}{n}\Sigma y(x - yCov(xy))(z - \gamma Cov(zy))$$

$$= Cov(xz) + \gamma^{2}Cov(xy)Cov(zy) + \frac{\gamma}{n}\ \Sigma xyz - 2\gamma^{2}Cov(xy)Cov(zy)$$

$$= Cov\ (xz)\left[1 + \frac{\gamma}{n}\ \frac{\Sigma xyz}{Cov(xz)} - \gamma^{2}\ \frac{Cov(xy)Cov(zy)}{Cov(xz)}\right]$$

$$= Cov(xz)\left[1 + D_{x}^{*} + D_{z}^{*} - \gamma^{2}V(y)\ \frac{Cov(xy)}{\sigma_{y}\sigma_{x}}\ \frac{Cov(zy)}{\sigma_{y}\sigma_{z}}\ \frac{\sigma_{x}\sigma_{z}}{Cov(xz)}\right]$$

$$= Cov\ \ xz\left[1 + D_{x}^{*} + D_{z}^{*} - \gamma^{2}V(y)\frac{r_{xy}\ r_{zy}}{r_{xz}}\right].$$

The true β^{*} may be written in terms of population moments:

$$14)\quad \beta^{*} = \beta - \alpha^{*}\phi = \frac{\beta - \alpha\phi}{1 - R_{xz}^{2}}$$

$$15)\quad \beta^{*} = \beta\ \frac{\left[1 - \frac{Cov(zy)}{V(z)}\ \frac{V(x)}{Cov(xy)}\ \frac{Cov(xz)}{V(x)}\right]}{1 - R_{xz}^{2}} = \beta\ \frac{\left[1-R_{xz}^{2}\ \frac{V(x)}{Cov(xz)}\ \frac{Cov(zy)}{Cov(xy)}\right]}{1 - R_{xz}^{2}}$$

$$16)\quad = \beta\ \frac{\left[1 - R_{xz}^{2}\ \ \frac{r_{zy}}{r_{xz}r_{xy}}\right]}{1 - R_{xz}^{2}} = \frac{r_{xy} - r_{xz}\ r_{zy}}{(1-R_{xz}^{2})r_{xy}}$$

$$17)\quad E(\beta_s^*) = E(\beta_s)\frac{\left[1\left(-\dfrac{\mathrm{Cov}(zy)\ \mathrm{Cov}(xz)\ [1 + D_z - \gamma V(y)]\left[1 + D_x^* + D_z^* - \gamma^2 V(y)\dfrac{r_{xy} r_{zy}}{r_{xz}}\right]}{V(z)\ \mathrm{Cov}(xy)\ [1 + D_x - \gamma V(y)]\ [1 + D_z - \gamma^2 V(y) R_{zy}^2]}\right)\right]}{1 - \dfrac{\mathrm{Cov}(xz)^2}{V(x)V(z)}\dfrac{\left[1 + D_x^* + D_z^* - \gamma^2 V(y)\dfrac{r_{xy} r_{zy}}{r_{xz}}\right]^2}{(1 + D_z^* - \gamma^2 V(y) R_{zy}^2)(1 + D_x^* - \gamma^2 V(y) R_{xy}^2)}}.$$

We assume that z and x are symmetric:

$$18)\quad E(\beta_s^*) = \beta \cdot Q_x \frac{\left[1 - \dfrac{R_{xz}^2\ V(x)\ \mathrm{Cov}(zy)}{\mathrm{Cov}(xz)\ \mathrm{Cov}(xy)}\ S\right]}{1 - R_{xz}^2 \cdot S \cdot T}$$

$$19)\quad = \beta \cdot Q_x \frac{\left[1 - R_{xz}^2 \dfrac{r_{zy}}{r_{xz}\ r_{xy}}\ S\right]}{\left[1 - R_{xz}^2 \cdot S \cdot T\right]},$$

where $Q_x = \dfrac{1 - \gamma^2 V(y)}{1 - \gamma^2 V(y)\ R_{xy}^2} < 1$ is the attenuation ratio for the first order regression

$$S = \frac{1 - \gamma^2 V(y)\ \dfrac{r_{xy}\ r_{zy}}{r_{xz}}}{1 - \gamma^2 V(y) R_{zy}^2} = \frac{r_{xz} - d^2 r_{xy} r_{zy}}{r_{xz}(1 - d^2 R_{zy}^2)}$$

$$T = \frac{1 - \gamma^2 V(y)\ \dfrac{r_{xy}\ r_{zy}}{r_{xz}}}{1 - \gamma^2 V(y)\ R_{xy}^2} = \frac{r_{xz} - d^2 r_{xy} r_{zy}}{r_{xz}(1 - d^2 R_{xy}^2)},$$

where $d^2 = \gamma^2 V(y)$.

We get the following reduction from (19) dividing by β^*:

$$20) \quad \frac{E(\beta^* s)}{\beta^*} = Q_x \frac{\beta}{\beta^*}\left(\frac{1 - SR^2_{xz}\dfrac{r_{zy}}{r_{xz}r_{xy}}}{1 - STR^2_{xz}}\right)$$

$$21) \quad = \left[\frac{1-d^2}{1-d^2R^2_{xy}}\right]\left[\frac{(1 - R^2_{xz})r_{xy}}{r_{xy} - r_{xz}r_{zy}}\right]\left[\frac{1 - \left(\dfrac{r_{xz} - d^2r_{xy}r_{zy}}{r_{xz}(1 - d^2R^2_{zy})}\right)\left(\dfrac{r_{xz}r_{zy}}{r_{xy}}\right)}{1 - \left(\dfrac{r_{xz} - d^2r_{xy}r_{zy}}{r_{xz}(1 - d^2R^2_{zy})}\right)\left(\dfrac{r_{xz} - d^2r_{xy}r_{zy}}{r_{xz}(1 - d^2R_{xy})}\right)R^2_{xz}}\right].$$

Cancelling r_{xz}'s in the third term of this product, and rearranging, we get

$$\frac{E(\beta^* s)}{\beta^*} = \frac{(1 - d^2)}{(1 - d^2R^2_{xy})}\left[\frac{\left(1 - R^2_{xz}\right)r_{xy}}{r_{xy} - r_{xz}r_{zy}}\right]$$

$$\left[\frac{[r_{xy}(1 - d^2R^2_{zy}) \qquad - r_{zy}(r_{xz} - d^2r_{xy}r_{zy})]\ (1 - d^2R^2_{zy})(1 - d^2R^2_{xy})]}{[(1 - d^2R^2_{zy})(1 - d^2R^2_{xy}) - (r_{xz} - d^2r_{xy}r_{zy})^2]r_{xy}(1 - d^2R^2_{zy})}\right].$$

Cancelling $(1 - d^2R^2_{xy})(1 - d^2R^2_{zy})r_{xy}$ from numerator and denominator, we get

$$= \frac{(1 - d^2)(1 - R^2_{xz})[r_{xy}(1 - d^2R^2{}_{zy}) - r_{zy}(r_{xz} - d^2r_{xy}r_{zy})]}{(r_{xy} - r_{xz}r_{zy})[(1 - d^2R^2_{zy})(1 - d^2R^2_{zy}) - (r_{xz} - d^2r_{xy}r_{zy})^2]}.$$

Computing the last term in numerator and denominator, we get

$$= \frac{(1 - d^2)(1 - R^2_{xz})[r_{xy} - d^2R^2_{zy}r_{xy} - r_{zy}r_{xz} + d^2R^2_{zy}r_{xy}]}{(r_{xy} - r_{xz}r_{zy})[1 - d^2R^2_{zy} - d^2R^2_{xy} + d^4R^2_{zy}R^2_{xy} - R^2_{xz} + 2d^2r_{xy}r_{zy}r_{xz} - d^4R^2_{xy}R^2_{zy}]}$$

$$= \frac{(1 - d^2)(1 - R^2_{xz})(r_{xy} - r_{zy}r_{xz})}{(r_{xy} - r_{xz}r_{zy})[1 - R^2_{xz} - d^2(R^2_{xy} + R^2_{zy} - 2r_{xy}r_{zy}r_{xz})]}$$

$$= \frac{(1 - d^2)(1 - R^2_{xz})}{1 - R^2_{xz} - d^2(R^2_{xy} + R^2_{zy} - 2r_{xy}r_{zy}r_{xz})}$$

$$22) \quad \frac{E(\beta^*_s)}{\beta^*} = \frac{1 - d^2}{1 - d^2 \dfrac{R^2_{xy} + R^2_{zy} - 2r_{xy}r_{xz}r_{zy}}{1 - R^2_{xz}}} = \frac{1 - d^2}{1 - d^2R^2} ,$$

where R^2 is the coefficient of determination of the regression (1) predicting y with x and z in the full population. The equivalence represented by (22) is proved in Johnston (1963, p. 57).

Thus, when both right hand side variables are symmetric and selection probabilities are a linear function of the dependent variable, the selection bias in partial regression coefficients is (1) the same in both variables;[2] (2) given by the same formula as for zero order regression coefficients; (3) smaller the larger the R^2 of the true multi-variate relationship; (4) smaller the smaller is the degree of selection on the dependent variable.

NOTES

[1] Homoskedasticity and the independence of x and u makes it possible to simplify $\Sigma y^2 x$ and Σyx^2:

$$\Sigma y^2 x = \Sigma\ x(\beta x + u)^2 = \Sigma\ \beta^2 x^3 + 2\beta x^2 u + \Sigma xu^2 = \beta^2 \Sigma x^3$$

$$\Sigma yx^2 = \Sigma\ x^2(\beta x + u) = \beta\Sigma x^3 + \Sigma x^2 u = \beta\Sigma x^3.$$

[2] In recent, as yet unpublished work, Arthur Goldberger (1975) has proved a result that is in many ways more general. When the right hand side variables are multi-normally distributed, truncation or selection bias results in a proportionate shrinkage of all regression slopes by $\theta^2/1 - (1 - \theta^2)R^2$, where θ^2 is the ratio of the restricted sample variance of y to the population variance of y. Note that $(1 - \theta^2)$ corresponds to $\gamma^2 V(y)$ in our notation. Thus, for the special case of bivariate and trivariate regressions when there is a linear relation between y and the probability of selection, this paper generalizes Goldberger's result to symmetric right hand side variables.

REFERENCES

Bishop, John. 1974. The private demand for places in higher education. Ph.D. dissertation, University of Michigan. Available from University Microfilms.

Cain, Glen. 1975. The challenge of dual and radical theories of labor market to orthodox theory. Discussion Paper 255-75. Institute for Research on Poverty, University of Wisconsin-Madison.

Crawford, David. 1975. Estimating earnings functions from truncated samples. Discussion Paper 287-75. Institute for Research on Poverty, University of Wisconsin-Madison.

Goldberger, Arthur. 1975. Linear regression in truncated samples. Unpublished manuscript.

Harrison, Bennett. 1972. Education and underemployment in the urban ghetto. American Economic Review 62:796-812.

Hausman, Jerry, and Wise, David. 1977. Social experimentation truncated distributions, and efficient estimation. Econometrica 45:919-938.

Johnston, J. 1963. Econometric methods. New York: McGraw-Hill.

Kmenta, Jon. 1971. Elements of econometrics. New York: MacMillan.

Manski, Charles, and Lerman, Steven. 1976. The estimation of choice probabilities from choice based samples. Unpublished paper. School of Urban and Public Affairs, Carnegie-Mellon University.

Masters, Stanley, and Ribich, Thomas. 1972. Schooling and earnings of low achievers: comment. American Economic Review 82:755.

Porter, Richard D. 1973. On the use of survey sample weights in the linear model. Annals of Economic and Social Measurement 2:141-158.

Taubman, Paul, and Wales, Terence. 1974. Higher education and earnings. New York: McGraw-Hill.