NBER Working Paper Series

ANALYSIS OF LONGITUDINAL EARNINGS DATA:
AMERICAN SCIENTISTS 1960-70

Lee A. Lillard
and
Yoram Weiss

Working Paper No. 121

CENTER FOR ECONOMIC ANALYSIS OF HUMAN BEHAVIOR
AND SOCIAL INSTITUTIONS

National Bureau of Economic Research, Inc.
204 Junipero Serra Boulevard, Stanford, CA 94305

January 1976

Preliminary; not for quotation.

NBER working papers are distributed informally and in limited number for comments only. They should not be quoted without written permission of the author.

This report has not undergone the review accorded official NBER publications; in particular, it has not yet been submitted for approval by the Board of Directors.

This research was supported by grants to NBER from the National Science Foundation (SOC71-03783) and from The Rockefeller Foundation (RF 72087). The opinions expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation or The Rockefeller Foundation. The authors wish to thank Barbara Williams for very competent assistance.

I. INTRODUCTION

Earnings functions which relate individual earnings to some, presumably relevant, characteristics are often used to predict the effect of individual choice on future earnings. Typically such estimates are based upon a single cross section. There are, however, well known difficulties in applying the results of such comparisons across individuals to the prediction of the earnings of any single individual.

One difficulty is that individuals of different ages who are observed at a given point of time may vary systematically with respect to their date of entry into the labor force. If either the slope or the level of individual profiles depends on vintage, then the cross section information is not sufficient to separate these effects from those associated with the accumulation of experience. Generally speaking, we expect the cross section to underestimate the effects of experience on earnings. This difficulty can be overcome only if we pool information from several successive cross sections. (See Weiss [11], and Welch [12]).

A second difficulty is that a snapshot at a given point of time provides no information on trends or any other dynamic changes which occur in the economy. We can use cross section data for prediction only in an economy which is either in a stationary or a steady state equilibrium. Again, by pooling information from several successive cross sections we can produce some estimates of the trends in earnings for various subgroups of the population.

A third difficulty is that individual earnings are also affected by some characteristics which the researcher is unable to observe. The strength of the longitudinal data is that it enables us to make use of the fact that

From Kelvin Smith - 2/18/76.

these unobservables are already to some degree incorporated in the past earnings of the same individual. A standard example is that of unmeasured ability. If the individual can be observed repeatedly one may incorporate the information on the existence of persistent effects (without actually measuring them) to improve the efficiency of the estimates of the earnings function. Furthermore it is important to know the relative importance of individual persistent effects relative to the total unexplained variance. We obtain from the model quite different distributional (welfare) implications if "errors" are uncorrelated over time. In such a case we would expect that over a whole life time the mean error is zero, and individuals with the same observable characteristics can be viewed as having identical "permanent" income. If however, there are persistent error elements an individual may be systematically below or above the mean of a group with similar observable characteristics.

Longitudinal data may also be useful in eliminating some selection biases which occur when different individuals are sampled. For instance, suppose that the longer the person is in the labor force, the more is learned (say by employers) about the individual's ability. Again this information on the individual, e.g., his professional reputation, is not known to the researcher. Suppose further that as a result of such learning individuals may be selected out of the sample (i.e., survival of the fittest), then it is clear that an estimate of experience effects which is based upon comparing different individuals (of the same vintage) may overestimate the potential gain for any single individual.

Finally, longitudinal data may be useful in reducing errors of measurement. One can measure with greater precision variables like schooling and experience which affect individual earnings. In this paper we present and compare estimates of earnings function based upon cross section and longitudinal data. Our source of data is the N.S.F. Register of Techinical and Scientific Personnel. This data is used to illustrate the main methodological issues.

The major findings of this study are as follows:

(1) Simple cross section estimates grossly underestimate cohort profiles during the period 1960-70. Furthermore the growth in earnings is not uniform across experience groups and more recent ~~earlier~~ vintages tend to have steeper profiles in most fields. Consequently the rate of return or present value comparisons based on cross sections are likely to be misleading even if the standard adjustment for growth is made.

(2) For purposes of estimating mean profiles and mean effects of variables estimates based on pooled independent cross sections are quite close to those based on the more expensive longitudinal data.

(3) There are important persistent unmeasured individual effects on both the level and growth of earnings. Consequently, individuals with the same observed characteristics will still have a wide variance in their permanent income.

II. THE EARNING FUNCTION

The earnings of an individual reflect current and past voluntary decisions and exogenous factors. An economic model is necessary to describe the precise nature of the interaction between individual choices and varying economic environments. In this paper we shall not attempt to provide such a model, (see however, Weiss [11]). We shall simply postulate the following simple model of earning determination.

$$(1) \qquad \ln Y_{it} = \ln Y_i^P + \ln Y_i^T .$$

The term $\ln Y_i^P$ denotes the systematic part of earnings (e.g., human capital) which reflects past decisions of the individual. The variable $\ln Y_i^T$ denotes the transitory element in earnings. It reflects decisions or exogenous events which affect only the current level of earnings without any future ramifications for individual earning capacity. The systematic part of earnings is determined by two linear relations which determine its initial level and growth:

$$(2) \qquad \ln Y_{i\mu}^P = \alpha_0 + X_i'\alpha_2 + V_\mu'\beta_1 + g\mu$$

$$(3) \qquad \frac{d\ln Y_{it}^P}{dt} = \alpha_1 + g + X_i'\alpha_3 + V_\mu'\beta_2 + Z_{it}'\gamma + T_t'\omega$$

where t is the year at which the individual is observed, and μ is the year at which the individual entered the labor force (we shall use year at which highest degree was obtained). The vector X_i denotes factors which are specific to the individual, remain fixed throughout his life, such as sex, level of degree and age at highest degree. The vector V_μ denotes factors which are common to individuals who entered the labor force in year μ, but may differ from year to year, such as the amount of knowledge provided by schools at year μ, size of entering cohort at μ and expectations formed at year μ. The vector Z_{it} contains variables which for every individual vary with time, such as his experience and type of employer. Finally the vector T_t denotes general market conditions at year t, as deviations from the general trend g.

The nature of our data is such that all individuals are observed over the same period of time. Starting salaries $\ln Y_{i\mu}$ are observed only for a minority of the sample. To estimate the parameters of both (2) and (3) it is necessary to integrate equation (3) and use information on current earning levels. For that purpose we shall make several highly restrictive assumptions:

(1) Apart from trend the individual expects no change in (real) market conditions. All departures from trend T_{jt} are viewed as unexpected and will enter the transitory part, $\ln Y_i^T$.

(2) The only individual variables varying with time are age and experience, which are linear functions of time. Type of employer is expected to remain fixed, and is included in the X vector together with the individual characteristics. Any deviations from the "normal" type of employment will be included in the transitory part, $\ln Y_i^T$.

(3) The effect of year of entry follows a simple growth trend, and can be indexed by the year of entry. We thus aggregate all the vintage effects into a single variable μ.

We can now integrate equation (3) to obtain a single level equation, which must hold at every point of time.

$$\text{(4)} \qquad \ln Y^{P}_{it} = \alpha_0 + \alpha_1(t-\mu) + X'_i(\alpha_2 + (t-\mu)\alpha_3) + \mu[\beta_1 + \beta_2(t-\mu)] + \frac{1}{2}\gamma_1(t-\mu)^2 + \frac{1}{2}\gamma_2(\tau^2_t - \tau^2_\mu) + gt$$

where μ is year of Ph.D., $(t-\mu)$ is experience, τ_t is current age, and τ_μ is age at attainment of Ph.D. Equation (4) is a straightforward generalization of the earnings function developed by Mincer [8]. The only difference is the explicit appearance of age at highest degree, age, vintage, and time as explanatory variables. These additions are suggested on a theoretical basis as well as by their availability in our data. However, an important problem of identification arises. There are two basic identities which may be noted.[1]

(5) $t = \mu + (t-\mu)$ or

current year of observation = year of highest degree + post degree experience.

(6) $\tau_t = \tau_\mu + (t-\mu)$ or

current age = age at highest degree + post degree experience.

It is therefore impossible to identify the separate effect of all these variables.

Furthermore, depending on the data we may have even further restrictions. If, for instance, we have only a cross section at a given point of time, the effect of t cannot be estimated at all, and experience cannot be separated from vintage. If we have time series data for a single individual, or a number of individuals of the same vintage, the effect of vintage cannot be estimated and we cannot separate time from experience (only $g + \alpha_1$ can be estimated). Since our data consists of time series on individuals of different vintages, by pooling them two of the effects can be identified in terms of the third. Specifically, we shall omit year of highest degree, age and the square of age from the estimation. The effects of these omitted variables are then captured by the coefficients of the remaining variables from each identity.

From equation (4) if μ is omitted, the coefficient of time will also capture the growth in starting salary which is vintage specific, β_1, in addition to the general growth in productivity, g, which accrues to all vintages. Similarly the coefficient of experience will be $(\alpha_1-\beta_1)$ and will be biased downward by the effect of vintage on the level of starting salaries. Finally, the coefficient of experience squared will be $(\gamma_1+\gamma_2)/2$ and will capture age effects. The effect of age on the curvature of the earning profile γ_2, will be identified by the coefficient of the interaction between age at highest degree and experience.

III. THE ERROR STRUCTURE

The purpose of this section is to describe the statistical procedures which will be used in the estimation of the earnings function. Two quite different models are considered. They are (1) the standard variance component method of pooling time series and cross section data, (See Wallace and Hussin [10], Maddala [7], and Nerlove [9]) and (2) a procedure which allows for first order serial correlation among individual observations. The variance component model implies the same correlation among all observations on the same individual but allows the correlation to vary among individuals. In the first order autoregressive model the correlation between observations on the same individual declines with their distance but is restricted to be the same for all individuals. The variance component model reflects the operation of unmeasured variables which vary among individuals but which do not vary during the decade. The autoregressive model reflects the aggregate effect of unmeasured variables which differ among individuals and among years for a given individual but which vary "smoothly" or not purely randomly over time. A comparison of the two estimates enables us to examine the sensitivity of coefficient estimates of the earnings function to the alternative specifications.

The simple variance component model, can be written as

$$(7) \qquad \mu_{it} = \delta_i + \varepsilon_{it} \quad . \qquad \begin{array}{l} i = 1,\ldots,N \text{ individuals} \\ t = 0,2,\ldots,10 \text{ time periods (two year} \\ \quad \text{intervals, 1960-1970).} \end{array}$$

where δ_i and ε_{it} are independent of each other and all included variables.

The δ component of this error structure reflects the effect of unmeasured variables in equation (2). It is thus a pure level effect. (In a later section we shall discuss the possibility of unmeasured variables which persistently affect earnings growth, equation (3)). Under this specification--the individual component δ_i represents a random variance component for which we need only estimate a mean and variance, rather than treating it as a parameter to be estimated.

For several reasons we have chosen to estimate time effects rather than treat them as random. First we feel they represent an important source of growth in earnings common to all individuals in the sample which should be measured even though it only measures our ignorance of the sources of that growth. Secondly, exogenous time effects are different for different subgroups of the sample, which would be hard to capture by a variance components procedure with respect to time effects. Thirdly and importantly, there are only six years in the NSF sample and more than six parameters are estimated so that far too few degrees of freedom are available for a time variance components procedure to be meaningful.

As is well known, the generalized least square estimate for this model is given by a weighted average of within individual and across individual, sample moments. Specifically

$$(8) \qquad \hat{\beta}_{GLS}=[\sum_{i=1}^{N} X_i' X_i - 6(1-\hat{\theta}) \sum_{i=1}^{N} \bar{X}_i \bar{X}_i']^{-1} [\sum_{i=1}^{N} X_i' \ell n Y_i - 6(1-\hat{\theta}) \sum_{j=1}^{N} \bar{X}_i \overline{\ell n Y}_i]$$

where X_i is the 6 x 1 data matrix for individual i, Y is annual earnings and θ may be estimated from the OLS residuals,[2] as $\hat{\theta} = \dfrac{\hat{\sigma}^2_\varepsilon}{\hat{\sigma}^2_\varepsilon + 6\hat{\sigma}^2_\delta}$.

The parameter θ represents the weight given to among versus within individual variation. A θ close to zero weights within individual variation, $(\ln Y_{it} - \overline{\ln Y_i})$, high relative to among individual variation $(\overline{\ln Y_i})$. That is, when θ is small due to θ^2_δ being large relative to σ^2_ε for a given T, a regression on means would be expected to have little predictive power relative to a regression on deviations since unobserved individual differences compose a larger share of the total residual variation. Conversely, when σ^2_δ is small relative to σ^2_ε for given T, a regression on means is expected to have greater predictive power relative to deviations and thus a large θ close to one is appropriate.

Usually alternative consistent estimates can be obtained by regression simply on means for individuals (using only among individual variation) or regression on deviations from individual means (using only within individual variation). A special characteristic of our model is that such regressions are not only less efficient but are also subject to difficulty in separating time and experience effects. Specifically, by using a regression on means across individuals the time effect cannot be identified and vintage effects cannot be separated from experience effects. Such a regression is analogous to a single cross section and is subject to the same problems. Similarly, by using only within individual variation the effect of vintage cannot be identified (as is true of any other variable whose effect is fixed throughout the period of observation) and the effect of time cannot be separated from the effect of experience. The parameter estimates represented by these two extreme procedures are not only orthogonal in the usual statistical sense (within group and between group estimates are always orthogonal) but are also orthogonal in their respective interpretations, level and growth, in the context of this model. Each is interesting in its own right. We are interested however in an incorporation of information from each source in an optimal way.

In Table 1 we present the various variance components by field. The second column gives the total residual variation due to all sources from the pooled OLS regression. The second column presents the individual level variance components. The proportion of the residual variation due to δ is 64 percent in the aggregate. That rather high proportion is remarkably stable across fields. The null hypothesis that $\rho = 0$ is strongly rejected.

In the case of a simple first order autoregressive residual structure among individual observations the error structure assumes the form

(9) $$\mu_{it} = \gamma\mu_{it-1} + \varepsilon_{it} \quad . \qquad \begin{array}{l} i = 1,\ldots N \\ t = 2,4,\ldots,10 \end{array}$$

It should be noted that the autocorrelation among residuals is net of time effects which are common to all individuals and are estimated directly. This model differs from the usual first order autoregressive model (see Durbin [2]) in two respects: (1) the residuals are blockwise autoregressive, within individual observations, and (2) observations are each two years apart. The generalizations are straightforward. In the procedure used here the parameter γ^2 is first estimated from the OLS residuals, R_{it}, by $R_{it} = \gamma^{\hat{2}} R_{t-2}$. Secondly the data matrix is tranformed by the matrix $I \otimes T$

Table 1. Variance Components by Field

Field	Sample size(6N)	$\hat{\sigma}^2_u$	Variance Components Assuming $\gamma = 0$					Serial correlation assuming $\sigma^2_\delta=0$
			$\hat{\sigma}^2_\delta$	$\hat{\sigma}^2_\varepsilon$	$\hat{\rho}$	$\hat{\theta}$	$1-\frac{\hat{\sigma}^2_\varepsilon}{\hat{\sigma}^2_{\ln Y}}$	γ
All Fields	67770	.0505	.0323	.0182	.639	.086	.802	
Chemistry	25980	.0510	.0325	.0185	.637	.087	.802	.883
Physics	9684	.0447	.0276	.0170	.619	.093	.827	.861
Biology	12960	.0568	.0363	.0205	.640	.086	.814	.857
Math	4548	.0502	.0341	.0161	.679	.073	.844	.880
Psychology	9816	.0535	.0351	.0184	.657	.080	.750	.885
Earth Sci.	3882	.0382	.0234	.0148	.614	.095	.798	.863

For description of data set, see p. 14 and pp. A1-A3.

$$(10) \qquad \text{where} \quad T = \begin{bmatrix} \sqrt{1-\hat{\gamma}^4} & 0 & 0 & 0 & 0 & 0 \\ -\hat{\gamma}^2 & 1 & 0 & 0 & 0 & 0 \\ 0 & -\hat{\gamma}^2 & 1 & 0 & 0 & 0 \\ 0 & 0 & -\hat{\gamma}^2 & 1 & 0 & 0 \\ 0 & 0 & 0 & -\hat{\gamma}^2 & 1 & 0 \\ 0 & 0 & 0 & 0 & -\hat{\gamma}^2 & 1 \end{bmatrix} \left(\frac{1-\gamma^2}{1-\gamma^4}\right)^{1/2}$$

Finally, OLS is applied to the transformed data.

Estimates of the parameter γ assuming no individual variance component ($\sigma_\delta^2 = 0$) are presented as the last column in Table 1. The values of γ are large and highly significant. Again there is considerable stability across fields.

Each of these two extreme models is superior to a purely stochastic specification and one can gain efficiency by using either of them. However, given the large number of observations in most fields there are only minor differences in the coefficients between the extreme variants of the GLS methods as well as between them and the OLS estimates (see Appendix Table A3 for the comparison of estimates for chemists). In the next section which analyzes the estimated earnings function we will use results based upon the simple variance components model.

We now turn to a more detailed analysis of the covariance structure across years for chemistry which is the largest single field. Consider first the correlation between log earnings in the various years. (Table 2.) There is a large positive correlation (76 percent to 87 percent) between successive years but the correlation declines as the observations become further apart.

Table 2. Log Earnings Variance and Correlation Matrix for Chemistry

	Variance	Correlation				
		1960	1962	1964	1966	1968
1960	.0936					
1962	.0851	.859				
1964	.0762	.817	.852			
1966	.0723	.741	.788	.871		
1968	.0768	.628	.677	.761	.820	
1970	.0830	.551	.606	.689	.758	.757

Table 3. Residual Variance and Correlation Matrix for Chemistry

	Variance	Correlation				
		1960	1962	1964	1966	1968
1960	.042					
1962	.045	.720				
1964	.044	.668	.741			
1966	.048	.590	.673	.800		
1968	.059	.498	.568	.682	.761	
1970	.068	.442	.517	.618	.703	.700

(The 1970-1960 correlation is only 55 percent.) In part, the positive correlation is due to measured individual variables which remain fixed throughout the sample period, such as sex, and type of employer. However, the correlation pattern of the residuals (Table 3) net of the various measured characteristics (including those which vary with time) is still of the same nature. There is a positive correlation (70 percent to 80 percent) between successive years which declines to 44 percent for observations which are a decade apart. This indicates the importance of persistent unmeasured effects. The question is whether they are best captured by a simple variance components model which assumes the same realization of the random unmeasured effect throughout the sample period, or by a simple autoregressive scheme which only requires that successive realizations be correlated.

The correlation matrix implied by the simple variance component model consists of a matrix with 1 on the diagonal and .637 for all off diagonal elements. The variance covariance matrix implied by the simple blockwise first order autoregressive model suggests a simple correlation of .884 among successive years (and correspondingly .781 among our observations which are two years apart) declining to .291 for observations a decade apart. The actual pattern of residual covariance declines monotonically with the distance between observations but not as rapidly as predicted by the simple autocorrelation model. It seems that a more general error structure incorporating both components would better fit the data.[3] However, since these two cases yield very similar estimates we will not proceed to develop the more general model.

IV. EMPIRICAL RESULTS: THE EFFECT OF TIME AND VINTAGE

The data for this study is a longitudinal file from the NSF Registry of Scientists. Each scientist in the sample reported at two year intervals, his earnings and personal and occupational characteristics. We focus on scientists with a Ph.D. degree. The scientists are classified into six fields on the basis of the field of greatest scientific competence. The fields are: Biology, Chemistry, Earth Science, Mathematics, Physics, and Psychology. In each field, as well as for the aggregate of all fields, we will estimate earnings function (4). More specifically, the following variables are included: The log of basic income as dependent variable; year of observation; type of employer including academic, government and private industry; quality of degree; the interaction between type of employer and year of observation; post degree experience; the difference between post degree experience and years since degree (=break); experience squared: and the interaction of experience with type of employer, age at highest degree, quality of degree, and break. In fields where there are a substantial number of women (more than 100) we have also added a female dummy and its interactions with experience. Detailed descriptions and summary statistics for the sample characteristics are presented in the Appendix.

It will be useful to introduce a more descriptive notation to be used hereafter and to rewrite equation (4):

$$\begin{aligned}
(11)\quad \text{log earning} &= a_{60} + b_{60}\ \text{private industry} + c_{60}\ \text{gov} + d_{60}\ \text{unstable} \\
&+ \text{yr}62\ (a_{62} + b_{62}\ \text{private industry} + c_{62}\ \text{gov} + d_{62}\ \text{unstable}) \\
&\ \ \vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots \\
&+ \text{yr}\ 70\ (a_{70} + b_{70}\ \text{private industry} + c_{70}\ \text{gov} + d_{70}\ \text{unstable})
\end{aligned}$$

$$+ \exp\,(\alpha_0 + \beta_1 \text{ private industry} + \beta_2 \text{ gov} + \beta_3 \text{ unstable})$$

$$+ \exp\,(\alpha_1 \text{ yhd} + \alpha_2 \exp)$$

$$+ \text{agehd } (\alpha_3 + \alpha_4 \exp)$$

$$+ \text{pre-experience } (\alpha_5 + \alpha_6 \exp)$$

$$+ \text{break } (\alpha_7 + \alpha_8 \exp)$$

$$+ \text{female } (\alpha_9 + \alpha_{10} \exp)$$

$$+ \text{top school } (\alpha_{11} + \alpha_{12} \exp)$$

The coefficients of this earning function estimated for each field and the aggregate of all fields by the variance component procedure outlined in the last section are presented in appendix Table 2.

Having estimated an earning function which incorporates time and vintage effects, we now are in a position to evaluate their separate roles in the determination of earnings. Specifically, we use equation (11) to produce two types of predictions:

1. Predicted cross section profiles, where year of observation is held constant but year of highest degree varies inversely with experience.
2. Predicted vintage profiles, where year of entry is held constant, but the year of observation varies together with experience.

Such predictions for all scientists in academic employment are presented in Figure 1.[4] There are two basic features of this diagram: (1) The level of

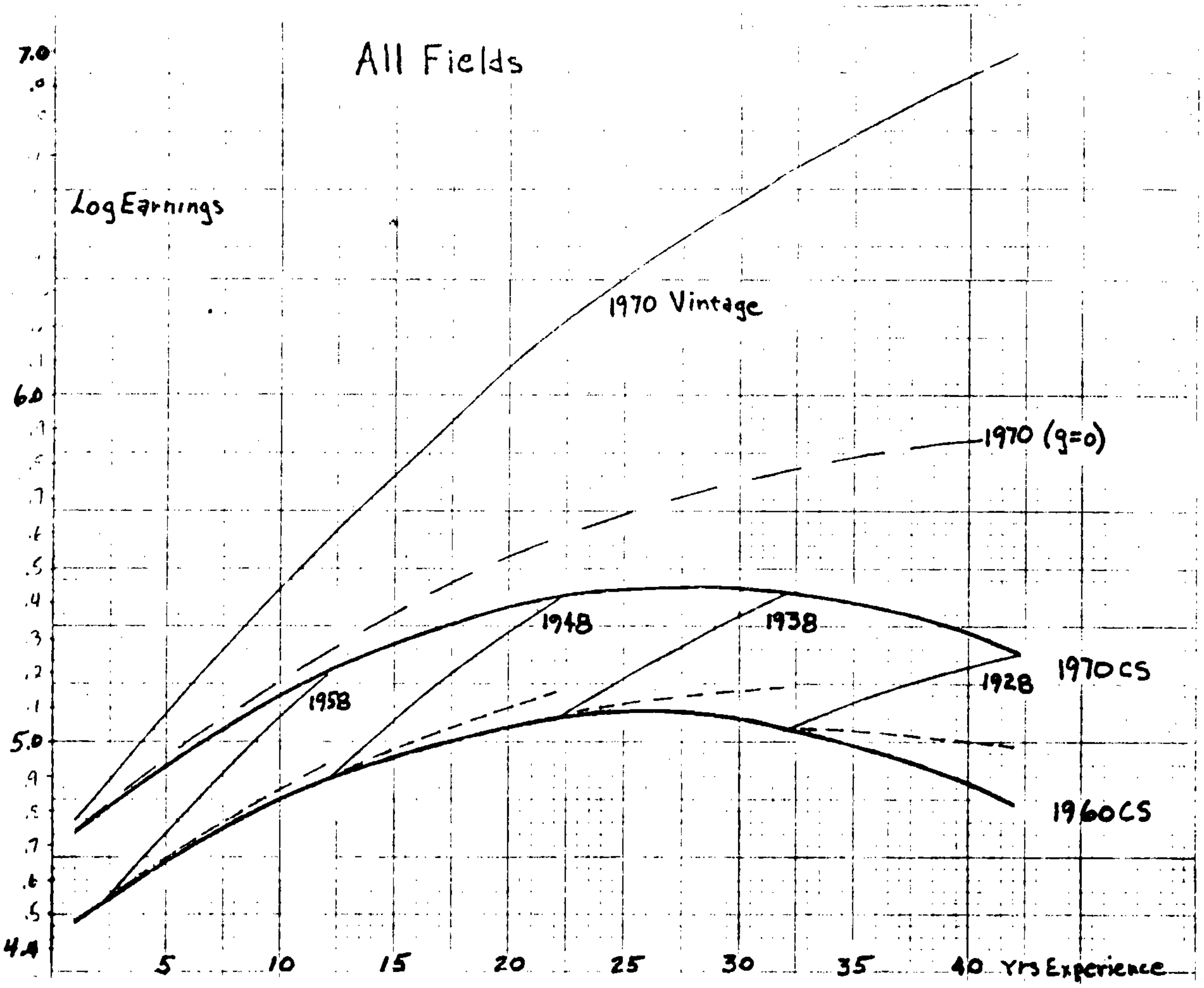


Figure 1. Log Earnings Profiles for the Aggregate of All Fields - Academics

Assumes Male = 1, Break in Experience = 0, Pre-Degree Experience = 0, Top Ten School = 0, Age at Highest Degree = 26.

later cross sections and vintage profiles is higher. (2) The slope of later cross sections and vintage profiles is higher. We shall discuss each of these effects separately.

It should be clear that our model does not allow us to separate the effect of time from that of vintage on the level of earnings. Thus the 1970 cross section may be higher than that of 1960, either because all vintages enjoyed the same general growth in productivity during the period, or alternatively, because more recent vintages start at a successively high initial salary. The interpretation of the cross section and vintage profiles will, however, be quite different. Under the first extreme interpretation the cross section profiles describe (in the absence of slope effects) the experience-earnings relationship. The vertical difference between cross sections reflects then the shift in individual lifetime earnings profiles due to growth in exogenous productivity.

Under the alternative extreme interpretation all growth in the level of productivity over time accrues to the entering cohort and experience earnings drift upward for successive vintages. The corresponding vintage profiles are the solid lines connecting the cross section profiles and labeled with the appropriate year of Ph.D. They represent the earnings paths actually attained on average. The contribution of experience for each vintage is given by the solid vintage line. The cross section profile then underestimates the experience effect by precisely the vertical shift.

The most likely situation is that both vintage and time effects are operating. Their effects can be separately identified only by analysis of the underlying causal factors influencing the productivity of individuals over time and of cohorts. There are, however, two weak indications that the shift in the level of the profiles is probably due to time effects rather than vintage. First the cross section profiles do not move smoothly through time. Individual year effects deviate significantly from the trend. Moreover, the individual deviations from trend correspond roughly to the changes in demand (and supply) conditions for scientists over this decade. (See Freeman [3].) We will use the term starting salaries to indicate the level effect since either the growth or the vintage interpretation has the same implication for starting salaries.

Let us now examine the changes in the slope of the earning profiles. The basic phenomenon is that holding experience constant, the slope of the individual earning profile is increasing with time (or vintage). This is also reflected by a higher vertical shift in the cross section profile for groups with more experience. There is thus a positive interaction between time or vintage and experience. In principle it is, again, impossible to separate time effects and vintage effects on the slope of the experience-earnings profile. Given the empirical results, either (1) during the 60's exogenous growth in productivity affected individuals with higher level of experience more favorably (a positive interaction between time and experience), or alternatively (2) more recent vintages tend to have greater earnings growth as well as a higher initial level (a positive interaction between year of Ph.D. and experience). One might argue that during the period under consideration, the time effects on slope would be different for different years depending upon the state of demand and supply.

We know that during the period conditions changed from rapid growth to a slow-down and in some cases reduction in real earnings toward the end of the period. It is likely that more recent and more experienced scientists fare differently under varying economic circumstances. Specifically one might expect that younger newly hired scientists will suffer more in the downturn and gain more in the upswing. If this were the case we would expect to find individual year effects on the slope which would be significantly different from a pure trend. This possibility was tested for two fields, physics, which underwent the sharpest changes in demand, and psychology. In each case there was no significant difference from trend. The finding of systematically greater earnings growth for scientists who are observed at a later date relative to those with the same years of experience but observed at an earlier date is probably due to their being of more recent vintage. A possible theoretical explanation is that the effect of more recent and higher quality schooling is not merely through the effect on starting salaries but also through a higher rate of investment on the job. (See Weiss [11]).

The broken lines, which are associated with each vintage in Figure 1 reflect estimates of the minimal effect of experience for these vintages.[5] They are calculated under the assumption of zero level effects for either time or vintage. The difference between the broken line and solid cross section line reflects the interaction between experience and vintage (or time). The difference between the broken line and the vintage profile reflects the contribution of time or vintage level effects. Even in the absence of level effects the cross section underestimates the experience profile of any given cohort since it also reflects the reduction in slope for older vintages. The bold lines for each vintage incorporate in addition to experience the level effects of time and vintage. The higher total growth in earnings during the sample period for more recent vintages (and lower experience groups) reflects the combined effects of non-linear individual profiles for each vintage, and the interaction between vintage and experience.

To highlight some of the problems associated with predicting individual earnings from cross section data,consider a person of 1948 vintage who was observed with 12 years of experience in 1960. Suppose we wish to predict his 1970 earnings. If we use the earnings of an individual with 22 years of experience in 1960, then we would underestimate his earnings by 35 percent. Even if there were no exogenous time effects on the level of earnings we would understate his earnings by 8 percent (the difference between the 1960 CS and the dashed line at 22 years experience).

The differences between cross section profiles and vintage profiles are highlighted most dramatically when we try to predict outside the sample period. Thus if one attempts to predict the lifetime earnings of a person who acquired his Ph.D. in 1970, the 1970 cross section provides a marked underestimate. This is true even if we assume that the growth in the level of earning which we observed during the 60's does not extend into the future (zero exogenous growth). This more conservative projection which allows vintage (or time) interaction with experience is given by the dashed 1970 profile. Clearly, if one assumes that on average the 60's trend coincides with the long run trend and that all growth in starting salaries was due to time, a considerably higher profile is indicated (the solid profile tagged 1970 vintage).

The shape of the earnings profile and the effects of time and vintage tend to vary from field to field and across type of employer. We shall not provide a detailed description of these differences. We shall only mention two extreme cases, physics and psychology. In psychology there is virtually no interaction between vintage (or time) and experience. Thus, apart from a correction for trend, cross section profiles provide an adequate description of the experience effects. In physics, there are relatively strong interactions between vintage (or time) and experience. Consequently, cross section profiles provide a misleading picture of the experience effects on earnings.

Comparison of Cross Section and Longitudinal Estimates.

Alternative predictions of cross section and vintage profiles based upon estimates from three different sources are presented in Figure 2. Included are generalized least squares estimates based upon longitudinal data on the same individuals, ordinary least square estimates based upon the pooling of three independent cross sections of different individuals, and cross section estimates in which separate regressions were fitted to independent 1960 and 1970 samples.[6]

Apart from relatively small differences the three estimates in Figure 2 are quite close.[7] This implies that for the purpose of predicting means conditional upon observed variables the relativelyless expensive independent samples are sufficient. These results also indicate that the earnings relationships is sufficiently stable over time to allow pooling of observations from different

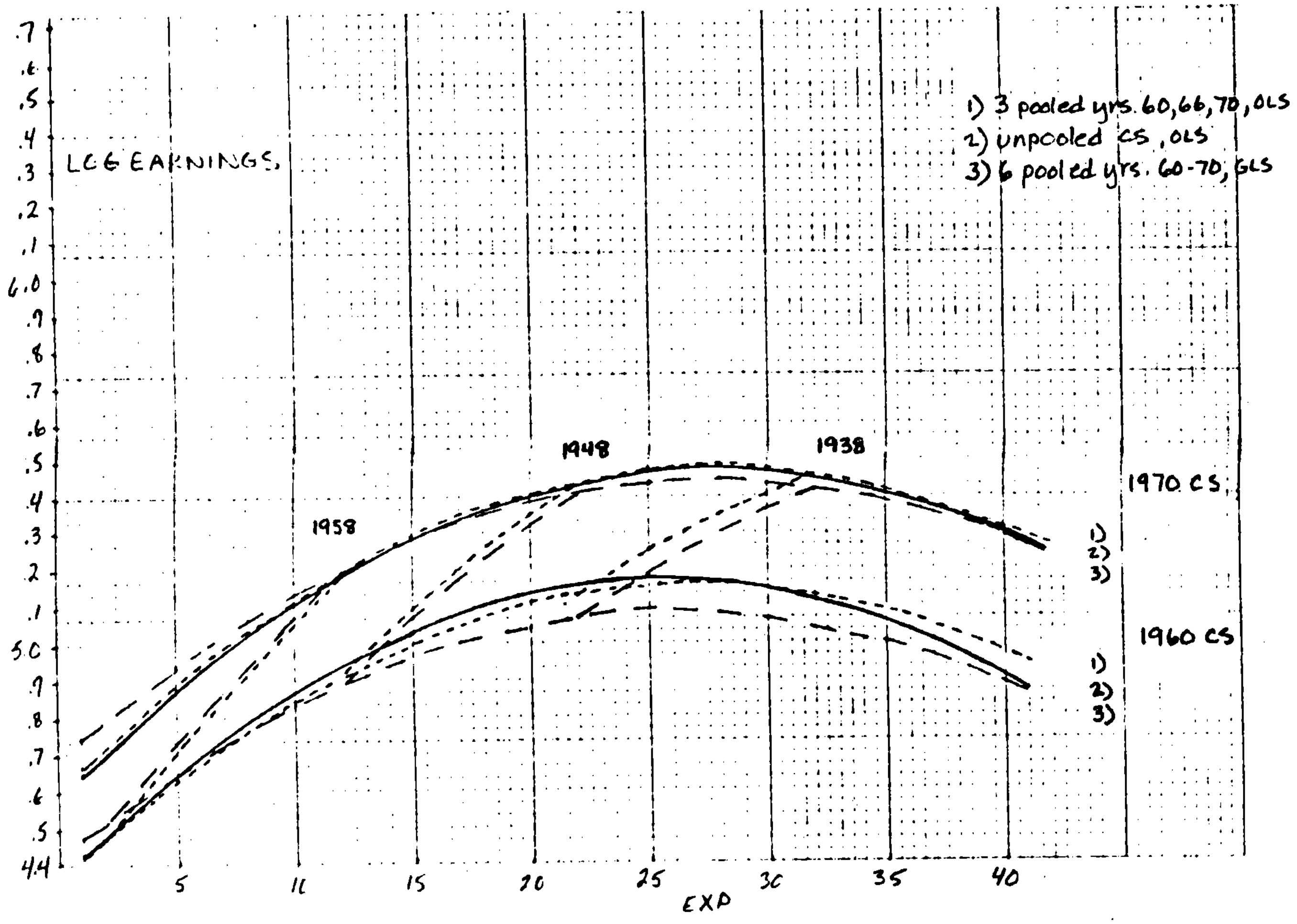


Figure 2. Comparison of Estimated Profiles by Data Source
All Fields - Academics

points in time without a complete set of interactions with time. The degree of closeness of fit, of course, is even higher within the same body of data (longitudinal or pooled independent cross sections). The close correspondence between the three estimation techniques (GLS, pooled OLS, separate OLS) is illustrated for physics in Appendix Figure 2. The fit is more remarkable considering the dynamic changes in demand for physicists which occured during the decade.

It is also worth noting that the GLS estimate is quite successful in predicting outside of the longitudinal age range (the separate cross sections each cover the full age range), e.g., at less than ten years of experience in 1970 and at older experience levels in 1960.

There are, however, some differences in the coefficients which are worth noting. (see Appendix Table 2, last two columns.) The coefficient of experience is considerably higher when estimated from pooled independent cross section than from longitudinal data (6.4 percent vs. 5.4 percent). To some extent this is compensated by a larger $\exp^2$ term, and the net result is that at low levels of experience higher contribution of experience is predicted. As we have already indicated this result may be interpreted as reflecting selection process whereby there is a greater probability of survival (in the Registry) among more able scientists which results in an overestimate of the experience effect for any given individual. This survival effect is absent in the longitudinal data since the same individuals are observed repeatedly.

Our ability to partially control selection is also useful in sharpening the precision of our estimate of the effect of the quality of degree. The quality of degree is defined here according to whether the school from which the Ph.D. degree was granted, was ranked among the top ten in the field. Using a pool of independent cross sections, there is again a selection bias, whereby the scientists of low ranking schools who survive in the sample are relatively more able, and thus the interactionbetween experience and quality of schooling is underestimated. In the longitudinal data we find a significant positive interaction between quality and the slope of the earning profile. (A similar result is reported by Johnson and Stafford [4] and [5]). This possibly indicates that more able individuals (who attend better schools) enjoy a higher "productivity" of experience. The absence of a significant negative difference in starting salaries indicates, as would be expected, that the quality of school difference in mean relative earnings is permanent and not compensated.

Another noticeable difference between the GLS and OLS estimates is the decrease in the coefficient of break. We define break to be the difference between years since degree and reported experience. To some extent this difference reflects genuine interruption in the accumulation of experience due, for instance, to military service. There is also the possibility of error in the measurement of experience. In the longitudinal data we assigned work experience to individuals who participated continuously on the basis of their average reported experience during the decade of observation. The result is a considerably smaller mean break and, more importantly, the variance in break is reduced. The GLS estimate indicates a negligible contribution to earnings of a year out of the labor force. This appears to be more plausible than the OLS estimate, which predicts that a person who is absent for a year, will upon his return enjoy a 4 percent increase in his wage, merely due to the increase in age. This growth rate is almost 70 percent of that of a scientist who participated continuously.

In the pooled independent cross section data, we estimated the effects of current type of employment on earnings. The longitudinal data allows us to obtain information of the normal employment of each individual. We classified the scientists in four groups: those who were continuously employed in the government in private employment and in academics, and a unstable group designating those scientists who changed employers at least once during the decade. (This group consists of 24 percent of the sample but includes non-reporters.) It is interesting to point out that the basic results are similar in the two samples. The scientists in private industry earn more than scientists in academics but this difference declines over time (33 percent difference in starting salaries in 1970 vs. 17 percent in 1970) and with experience. There is a negative interaction between employment in private industry and experience indicating that scientists in academics invest more in on-the-job-training. There are, however, significant differences between the OLS and GLS estimates of these coefficients. The longitudinal data estimates imply a sharper reduction in the advantage of private industry over time, and a weaker negative interaction of experience with employment in private industry.

V. TRANSITORY AND PERMANENT EARNINGS DIFFERENTIALS

In this section we provide an analysis of the variance in earnings focusing on the following issues: (1) the portion of the variation in earnings which can be attributed to differences in measured characteristics, and the portion due to unmeasured individual characteristics; (2) the portion of income variation at a given experience level that is permanent, and the part which reflects voluntary investment decisions, and is thereby compensated at later stages of the scientist's career. (3) the share of purely random and transitory effects.

In Figure 3 we present the frequency distribution of scientists' log earnings in 1970. The distribution has a mean of 5.35 (this corresponds to a level of about 21,000 dollars per year). About 4.4 percent of the scientists earn more than 50 percent above the mean (i.e., above 5.85) and only 3.1 percent earn less than 50 percent below the mean (i.e., below 4.85). Inequality as measured by the variance of the log of earnings, among individuals in 1970 was .081. In comparison the variance in the log of earnings due to observed characteristics is .017 (21.0 percent) and the variance in the individual δ component is .032 (39.5 percent) which means that the unobserved characteristics are more important than observed characteristics in explaining earnings inequality. This result is quite stable across fields.

Conceptually δ captures the proportional effect of individual ability, family, and social background and other unobserved differences which remain fixed throughout the period. What we identify of course is only the net effect of these factors. It is important to note these effects do not cancel

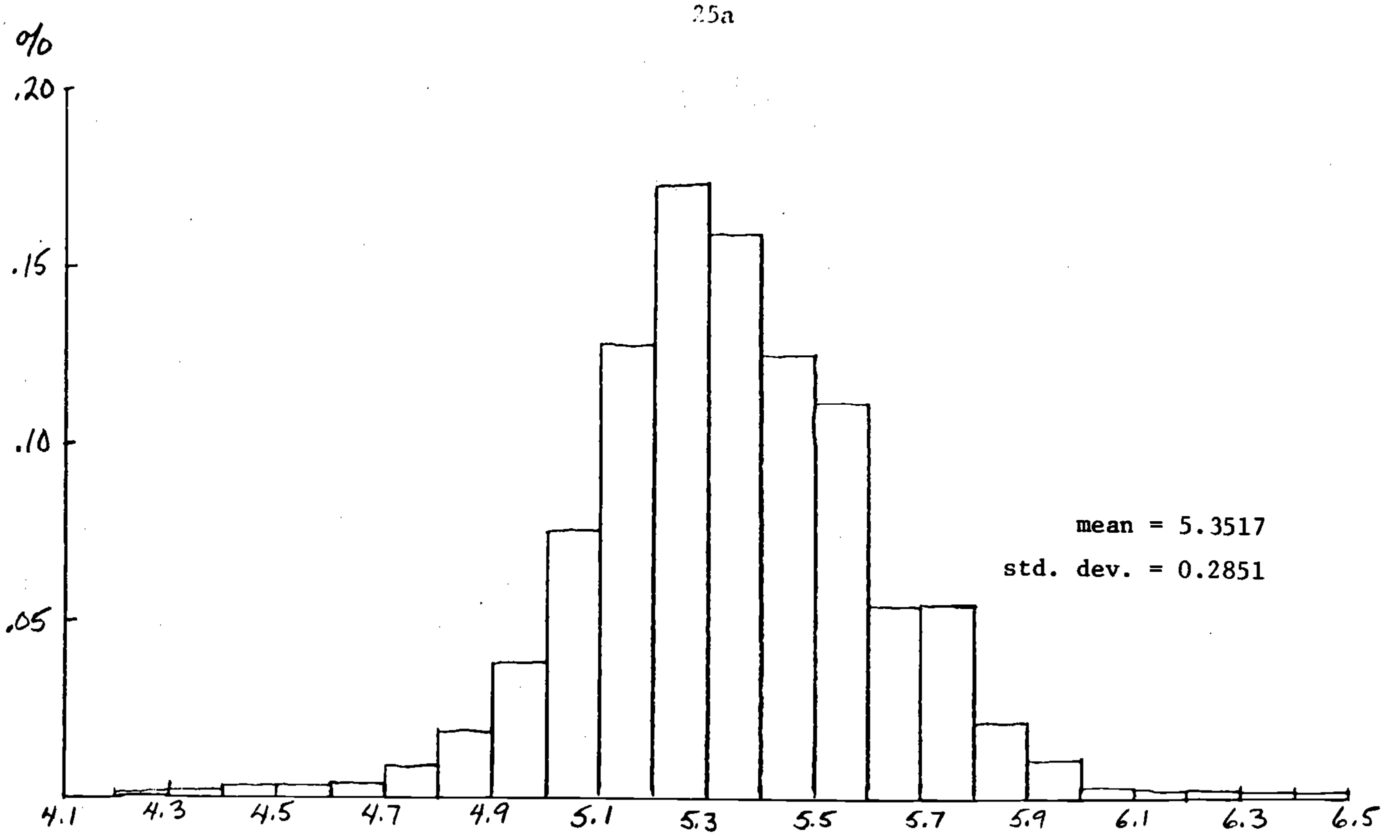


Figure 3. Distribution of 1970 Log Earnings.

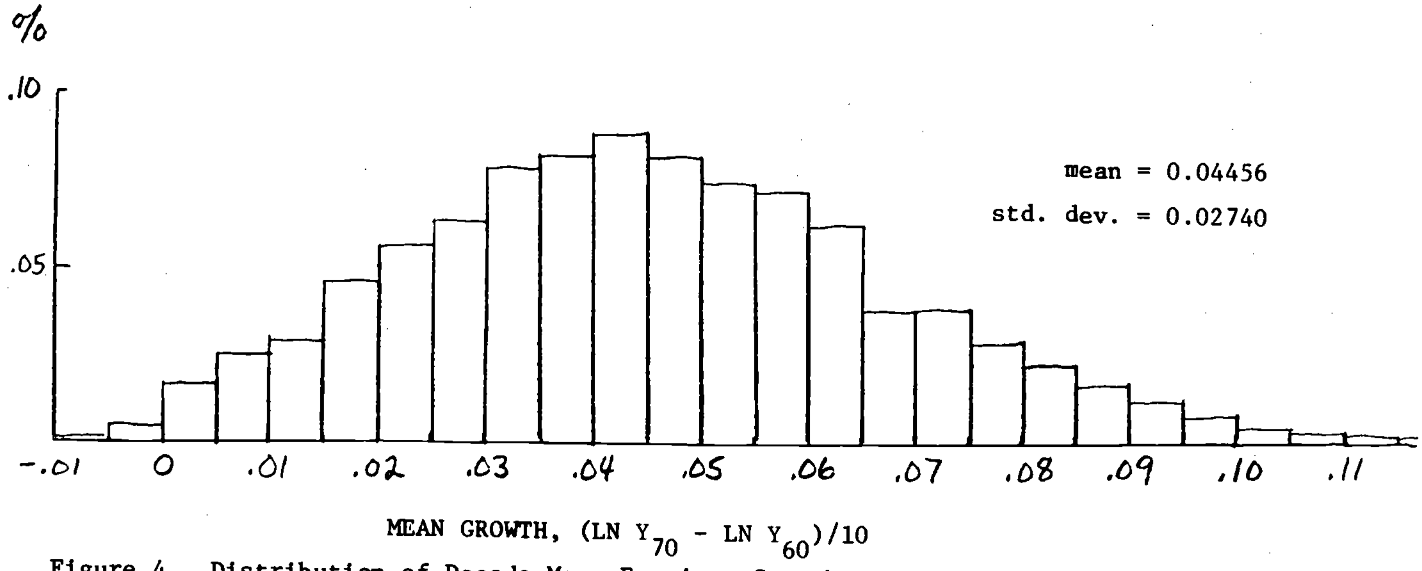


Figure 4. Distribution of Decade Mean Earnings Growth

each other out on average. Therefore δ reflects an unobserved permanent income component. Thus $\hat{\sigma}^2_\delta$ is a measure of permanent income inequality among observationally alike individuals at a point in time.

The effect of the transitory element ε, is the remainder after accounting for the variance of δ and the explanatory variables. This remainder is about 35 percent of the total variation, for the aggregate of all fields. This source of variation should be excluded from any measure of permanent earnings inequality. Note however that the differences $(\hat{\sigma}^2_{\ln Y_{70}} - \hat{\sigma}^2_\varepsilon)$ is only an upper bound on permanent earnings inequality in 1970 since it includes variation due to differences in experience which may be compensated over a lifetime.

Some insight into the role of transitory and experience related effects can be gained by examining the distribution of the mean log earnings over the decade. The distribution of these means can be viewed as a rough proxy for life time earnings.[8] Since the data covers only 10 years, its variance will overstate inequality in human wealth. The variance in means is .067 which is greater than 1/6 of the variance in any single year. (It is 82 percent of the 1970 variance, rather than 17 percent.) This reflects the fact that while part of the experience related differences among individuals are eliminated there remains a very substantial permanent component (see Lillard [6] for a further analysis of this issue).

The longitudinal nature of our data provides an opportunity to examine size distribution of earnings growth rates. The relative frequency histogram for mean real annual growth during the decade 1960-70 over individuals is presented in Figure 4. The distribution has mean .0466. On the average, the real earnings of scientists during the decade grew by almost 5 percent per year. There is also a substantial amount of variation around this mean; the standard deviation is .0274. Approximately 2 percent of the sample actually sustained a negative ten year mean growth rate. In the other extreme the proportion of scientists who enjoyed a real annual rate of above 10 percent is 2.4 percent. The effect of measured variables on the growth in earnings is approximately the same as their effect on the level. The aggregate variance in the mean annual growth rate during the decade was .00075. In comparison the variance due to observed individual variables is .00017, (22.7 percent) while the variance in persistent individual residual growth rate differences is .00029 (38.7 percent). There thus remains an unexplained 38.7 percent. Once again, these shares are quite stable across fields (see Table 4).

The approximate normality of both the level and growth distributions is roughly consistent with a simple random walk model for the log of earnings (see Aitchison and Brown [1]). Our data, as well as theoretical considerations suggest, however, that such a model is an oversimplification. Consider, for instance, the substantial negative correlation between the growth in earnings and the mean level of earnings over the decade. (See Table 4.) To some extent this correlation reflects a life cycle phenomenon, whereby earning is positively related and growth is negatively related to experience. This concavity property of the log earning profiles is a widely documented empirical phenomenon and is consistent with the findings reported in the last section. A more interesting question is the simple correlation between the level and growth in earnings net of the effect of measured variables.

Table 4. Residual Analysis by Field.

FIELD	$\hat{\sigma}^2_{(\mathrm{LnY}_{70}-\mathrm{LnY}_{60})/10}$	$\hat{\sigma}^2_{\xi}$	$\dfrac{\hat{\sigma}^2_{\xi}}{\hat{\sigma}^2_{(\mathrm{LnY}_{70}-\mathrm{LnY}_{60})/10}}$	CORREL $\overline{\mathrm{LnY}}$ $(\mathrm{LnY}_{70}-\mathrm{LnY}_{60})/10$	CORREL $\hat{\delta}$, $\hat{\xi}$
All Fields	.00075	.00029	.386	-.1259	.138
Chemistry	.00080	.00031	.388	-.1067	.220
Physics	.00072	.00028	.389	-.2056	.059
Biology	.00080	.00029	.363	-.1621	.058
Math	.00067	.00027	.403	-.2038	.007
Psychology	.00074	.00031	.419	-.0179	.126
Earth Science	.00058	.00019	.326	-.1447	.235

Only the omitted variables which affect the level of earnings have been accounted for by the simple GLS procedure which we adopted. Theoretical considerations underlying Equations (2) and (3) suggest that a persistent individual element is likely to occur in the growth equation as well. Put differently there is a possibility of interaction between experience or time and individual unobserved characteristics. (For a more detailed discussion of these and related topics, including the role of ability and access to capital markets see Lillard [6] and Weiss [11].)

We may utilize the longitudinal nature of the data to analyze the joint distribution of individual level and growth components of residual variation. The residual structure may be reformulated as

$$(12) \qquad \mu_{it} = \delta^*_i + \xi_i \ell + \emptyset_{it} \qquad \ell = -5,-3,-1,1,3,5$$

The parameter ξ_i represents the effect of omitted variables which are individual specific but which alter the earnings-experience or earnings-time relationship. The term $\xi_i \ell$ represents the aggregate effects of both $(t-\bar{t})$ and $(\mathrm{Exp}_{it} - \overline{\mathrm{Exp}_i})$ which cannot be separately identified for the reasons outoined earlier. The δ^* term correspondingly takes new meaning under this model, i.e., $\delta^*_i = \delta'_i + \xi_{i\ \mathrm{time}} \cdot \bar{t} + \xi_{i\ \mathrm{Exp}} \cdot \overline{\mathrm{Exp}}$. Since it is not unreasonable to expect some of the same unobserved variables to affect both δ^* and ξ, it is an empirical question whether the two are correlated. A positive correlation between ξ and δ^* implies that individual profiles are diverging from the predicted profiles over a life time. A person whose observed earnings at a given point of time exceed, for example, the mean earnings of observationally identical individuals, is not only likely to maintain this discrepancy on the average, but also to increase it. Strictly speaking the correlation pattern between δ^* and ξ_i relates only to whether unobserved individual effects are compensated during the decade of observation. We can justify a life time interpretation only if $\xi_{i\ exp}$ is "large" relative to $\xi_{i\ time}$.

As we have already indicated the simple correlation between mean relative earnings lnY and mean decade growth rate $(\ln Y_{70} - \ln Y_{60})/10$ is consistently negative and in the 10 to 20 percent range. In contrast the simple correlation between δ^*_1 and ξ_i (net of all measured year, vintage and experience effects) is consistently positive (see Table 4). The strength of the relationship varies from negligible in math to nearly 24 percent in earth science. The 22 percent estimate for chemistry is especially significant due to the large sample size. This can be interpreted as indicating an overlap of variables affecting both the level and growth of earnings in the same direction. This is weak evidence that the effect of these variables is not compensated in the sense of offsetting lower relative earnings early in the life cycle with greater relative earnings later.

Since we observe every individual for only 10 years it is necessary to consider the evidence within narrower experience groups. It is possible that differences which are not compensated during the decade are nevertheless compensated over the life cycle. In this case one would observe a U-shaped pattern for $\sigma^2_{\delta^*}$ (see Mincer [8]), and a correlation between δ^* and ξ which is negative at low levels of experience and positive later. If on the other hand the variance in δ^* increases with experience and level and growth individual components are positively correlated then unobserved profiles are diverging from the predicted profiles over the lifetime. An interesting special case is when individual level and growth are negatively correlated at early experience levels and positively correlated at late experience levels. In this case the actual lifetime profiles may cross the predicted ones. If lifetime profiles cross the predicted ones then the unobserved deviations are roughly compensated over a life time. If the profiles diverge the compensation does not occur which implies greater inequality among individual scientists than the one which is indicated by the variance in observed characteristics. Care must be taken in drawing inferences, however, since experience as of 1960 and vintage are linearly related.

The actual pattern of residuals across various experience groups are presented in Table 5 for chemists. Most fields are like chemistry except with weaker positive correlation between δ^* and ξ.[9] The tendency of σ^2_δ to increase with experience, and the positive correlation between ξ and δ^* indicates a mild "fanning out" of the relative earnings profiles from early to late experience levels. This pattern implies that the tendency is for unobserved differences in individual profiles not to be compensated over the life cycle. They do not then represent different investment patterns with the same lifetime earnings wealth. They indicate that those with greater relative earnings are also experiencing greater earnings growth, at all experience levels. This implies greater inequality in lifetime earnings wealth than is implied by the predicted mean profile presented earlier. Alternatively the pattern for chemistry may represent a greater dispersion in both mean relative earnings and mean growth rates for older vintages.

Table 5. Variance-Covariance of δ and ξ by Experience for Chemists.

YRS EXP 1960	SAMPLE SIZE	$\hat{\sigma}_u^2$	$\hat{\sigma}_\delta^2$	$\hat{\sigma}_\xi^2$	Corr$(\hat{\delta},\hat{\xi})$
Overall	4330	.05104	.03254	.00031	.220
0-5	1075	.02955	.01113	.00030	.329
5-10	1332	.03546	.01986	.00027	.238
10-15	800	.05295	.03620	.00034	.307
15-20	531	.07174	.05718	.00034	.137
20-25	402	.09857	.07297	.00029	.155
>25	190	.11482	.07209	.00038	.125

FOOTNOTES

[1] Strictly speaking, we should add in each identity a variable indicating possible breaks in the accumulation of experience. For simplicity we omit this variable from the analysis but will introduce it in the estimation.

[2] There are several ways to estimate θ from the data. Maddala suggests an analysis of covariance of the model allowing between individual (σ_{δ}^{2}) and within individual $(\sigma_{\varepsilon}^{2})$ effects. Nerlove [9] suggests the alternative of simply taking the variance of estimated dummy variables. In our case there are an intractable number of dummy variables to be estimated (always over 150 and often more than 2000). An alternative we consider is to estimate the dummy variables $\hat{\delta}_i$ in a two stage procedure. First estimate the model by OLS then estimate each individual parameter by the mean residual. Clearly these estimators $\hat{\delta}$, are unbiased. The corresponding variance of the estimator is not an unbiased estimator of the true variance. A correction is made for the sample variance of $\bar{\varepsilon}_i$ resulting from the relatively few yearly observations for each individual.

An alternative estimator can be derived from the residual variances in the following two regressions: (1) The regression of means over years corresponding to between individual variation. (2) The regression of deviation from individual means which corresponds to within individual variation. Again, a correction for the sample variance in ε_i, resulting from the fact that we have relatively few yearly observations, is necessary to obtain unbiased estimates. We have experimented with both methods of estimation of the residual variances, and found the differences in θ to be negligible.

The results of this comparison for Chemists are given in the Table below:

Alternative estimates of variance components based on 25,980 observations on 4330 chemists.

		Unbiased Estimates Of				
	Method	σ^2_ε	σ^2_δ	ρ	θ	$1-\theta$
(1)	Pooled OLS (Y_{it}) Residual Analysis	.01850	.03254	.637	.087	.913
(2)	Means Equation ($\bar{Y}_i$) & Deviations ($Y_{it}-\bar{Y}_i$)	.01833	.03271	.641	.085	.915

[3] We experimented with an error structure of the form

$$\mu_{it} = (1-\gamma)\delta_i + \gamma\mu_{it-1} + \varepsilon_{it} \qquad 0 < \gamma < 1.$$

This model incorporates both an individual variance component and serial correlation. The γ component represents first order serial correlation among individual observations net of the δ component $(\mu_{it} - \delta_i) = \gamma(\mu_{it-1} - \delta_i) + \varepsilon_{it}$ and is assumed the same for all individuals.

$$\rho = \frac{\sigma^2_\delta}{\sigma^2_\delta + \frac{1}{1-\gamma^2}\sigma^2_\varepsilon}$$

The variance-Covariance matrix for this model is

$$E(UU') = (\sigma^2_\delta + \frac{1}{1-\gamma^2}\sigma^2_\varepsilon)\ I \times A$$

$$\text{where } A = \begin{bmatrix} 1 & \rho+(1-\rho)\gamma & \rho+(1-\rho)\gamma^2 & \cdots\cdots & \rho+(1-\rho)\gamma^5 \\ & 1 & \rho+(1-\rho)\gamma & \cdots\cdots & \rho+(1-\rho)\gamma^4 \\ & & 1 & \ddots & \vdots \\ & \text{sym} & & \ddots & \vdots \\ & & & & 1 \end{bmatrix}$$

Obviously the special case of the simple individual variance component obtains when $\gamma = 0$ and the special case of blockwise serial correlation obtains when $\delta_i = 0$ and all of the forms developed in the text apply. This combined model should simulate the observed residual variance-covariance matrix more closely than either of the special cases. While we were unable to obtain consistent estimators for γ and ρ under the general model, ad hoc estimates derived by correcting for asymptotic bias reproduced the residual covariance quite well.

[4] In constructing these profiles we assumed age at highest degree = 4 (age is measured from 22), pre-experience = Break = Female = Top school = 0. The 1970 cross section profile is given by:

$$\begin{aligned} \ln Y = {} & a_{60} + a_{70} + \alpha_0 \exp + \exp\ (\alpha_1(1970-1958-\exp) + \alpha_2 \exp) \\ & + 4\ (\alpha_3 + \alpha_4 \exp) \end{aligned}$$

(Year of highest degree is measured from 1958).

The 1958 vintage profile is given by:

$$\ln Y = a_{60} - 2g + (\alpha_0 + g)\exp + \alpha_2 \exp^2 + 4(\alpha_3 + \alpha_4 \exp)$$

where $g = \frac{a_{70}}{10}$.

[5] The 1958 vintage profile, which assumes no growth, i.e. the broken line in figure 1, is given by

$$\ln Y = a_{60} + \alpha_0 \exp + \alpha_2 \exp^2 + 4(\alpha_3 + \alpha_4 \exp)$$

[6] Three independent random samples of 10,000 scientists each were drawn from the unmatched National Registers of 1960, 1966, and 1970. For further analysis of this data see Weiss [11].

[7] The relatively inferior correspondence in 1960 reflects the absence of data for 1960 on whether an academic scientist's earnings were for 9 months or 12 months. A correction was imputed from the data which is available in the other years. A different method was used in the longitudinal where information on the same individual in other years was used, and the independent cross sections where the sample mean probability of 9 month salary was imputed to all academics. Comparisons for 1966, in which the problem does not arise, are presented in appendix Figure 1. Like the 1970 comparison they support the conclusion of the text.

[8] These are geometric 10 year means which underestimate the arithmetic means. The latter is of course more relevant to present value calculations. Also discounting is ignored.

[9] The pattern for biology is one in which early in the life cycle there

is a positive correlation between level and slope effects and late in the life cycle they are negatively correlated. This, along with the increase in the variance of δ with increased experience, implies a "bow" in the individual lifetime profiles among individuals. Profiles diverge from the predicted ones early in the life cycle then converge to it. Alternatively there are substantial differences between younger and older vintages with the difference occurring at roughly 1940.

REFERENCES

[1] Aitchison, J. and J.A.C. Brown. The Lognormal Distribution, Cambridge University Press, New York, 1957.

[2] Durbin, J. "Estimation of Parameters in Time-Series Regression Models," Journal of Royal Statistical Society, Series B, Vol 22, 1960.

[3] Freeman, R. "Supply and Salary Adjustment to the Changing Market Science Manpower Market: Physics 1948-1973," American Economic Review LXV (March 1975), pp. 27-39.

[4] Johnson, G. and F. Stafford, "Lifetime Earnings in a Professional Labor Market: Academic Economists," Journal of Political Economy, 82, (May/June 1974), pp. 549-570.

[5] ______________________, "The Earning and Promotion of Women Faculty," American Economic Review, LXIV, (December 1974), pp. 888-903.

[6] Lillard, L. "Inequality: Earnings versus Human Wealth," forthcoming American Economic Review

[7] Madalla, G.S. "The Use of Variance Components Models in Pooling Cross Section and Time Series Data," Econometrica, Vol.39, No. 2 (March 1971), pp. 341-357.

[8] Mincer, J. Schooling Experience and Earnings, New York, National Bureau of Economic Research, 1974.

[9] Nerlove, M. "A Note on Error Components Models," Econometrica, Vol. 39, No. 2 (March 1971), pp. 383-396.

[10] Wallace, T.D. and A. Hussain. "The Use of Error Components Models in Combining Cross section with Time Series Data," Econometrica, Vol. 37, No. 1, (January, 1969), pp. 55-72.

[11] Weiss, Y. "The Earnings of Scientists, 1960-1970: Experience, Age, and Vintage Effects," NBER Working Paper #99, 1975.

[12] Welch, F. "Black/White Differences in Returns to Schooling," American Economic Review LXIII (December 1973), pp. 893-907.

APPENDIX

Description of Variables

YR62, YR64, YR66, YR68, YR70 = 5 year dummies. YR1960 is the omitted one.

PRE EXP = pre degree experience, professional work experience prior to obtaining highest degree.

TOPSC = Top ten school dummy. The rank of each scientist's university was established upon rankings provided by Cartter for 1964 and 1969, Keniston for 1957 and Hughes for 1925. See Johnson and Stafford (1974). Those scientists who attended a school in the top ten percent of this ranking were assigned this dummy variable.

BRK = Break or interruption in post degree experience. It is the difference between years since degree and post degree experience.

AGEHD = Age at highest degree.

EXP = Post degree professional experience, based on an average 1960 experience level. If experience was reported for any of the 6 years it was translated to 1960 experience and averaged over the reported years.

YHD = Year of highest degree.

FEMALE = Female sex dummy variable.

LOG OF EARNINGS = Dependent variable, log of basic salary in hundreds of 1970 real dollars. Academic scientists' salaries in 1962-1970 were adjusted to a 12 month basis if a 9 month salary was indicated. In 1960, no indicator was available so 1960 salaries were adjusted upward by a predetermined probability that their basic salary was reported on a 9 month basis.

Type of employer during the 10 year period:

PI = Private industry.

G = Government and non-profit institutions.

U = Unstable, i.e., the respondent switched type of employers over the 10 year period.

Academic employer is the omitted clas.

Description of NSF Data

1) Total number of observations = 104,906

2) Criteria for rejecting

   A) If highest degree is < B.A. or non-reported or a foreign degree or an associate.

   B) If (YHD - yr. of birth) < 15 yrs. for a B.A. or if (YHD - yr. of birth) < 17 yrs. for a M.A. or Ph.D.

   C) If YHD or yr. of birth not reported.

   D) If level of highest degree ≠ level of highest degree in 1970.

   E) If individual was any type of student in 1970.

   F) If major of highest degree ≠ first specialty in any reported year for which specialty is reported after YHD.

   G) If individual was a student for any reported year after YHD.

   H) If individual's employment status was not reported for every year after YHD.

   I) If both gross and basic income were never reported for all years after YHD and if he was not unemployed or not retired. (i.e., use if any income was reported after YHD even if he was unemployed or retired)

   J) If YHD ≥ 1960.

   K) If individual was not a fully employed non-student every year.

   L) If basic salary was not reported for every year.

   M) If gross salary was substituted for basic salary in any year.

Appendix Table 1. NSF Longitudinal Sample Characteristics

	Field						
	Aggregate	Chemistry	Physics	Biology	Math	Psych	Earth Sci.
Sample Size	11295	4330	1614	2160	758	1636	647
Type of Employer (%)							
Academic	.382	.208	.372	.552	.679	.446	.451
Government	.102	.053	.064	.138	.032	.185	.236
Industry	.274	.564	.205	.072	.074	.014	.130
Unstable	.237	.175	.358	.236	.215	.322	.182
Ph.D. at Top Ten Grad. School	.302	.244	.354	.333	.352	.290	.487
Year of Ph.D.							
Mean (1900s)	50.4	49.6	51.3	49.6	51.0	52.1	51.1
Std. Dev.	7.3	7.3	7.1	7.6	8.0	6.6	7.4
Age of Ph.D.							
Mean	29.3	28.1	28.8	29.3	29.6	31.5	30.9
Std. Dev.	4.2	3.3	3.6	4.2	4.7	5.4	4.4
Experience in 1960							
Mean	9.5	10.2	8.6	10.0	9.8	7.8	8.9
Std. Dev.	7.2	7.3	7.0	7.4	7.8	6.4	7.3
Pre-Degree Exp.							
Mean	2.6	1.9	2.6	2.2	3.5	3.7	4.2
Std. Dev.	3.5	2.8	3.3	3.4	4.3	4.3	4.0
Break in Exp.							
Mean	.17	.12	.08	.40	.16	.10	.07
Std. Dev.	.78	.67	.56	1.1	.72	.77	.49

Appendix Table 2: Regression Estimates by Field and Estimation Technique

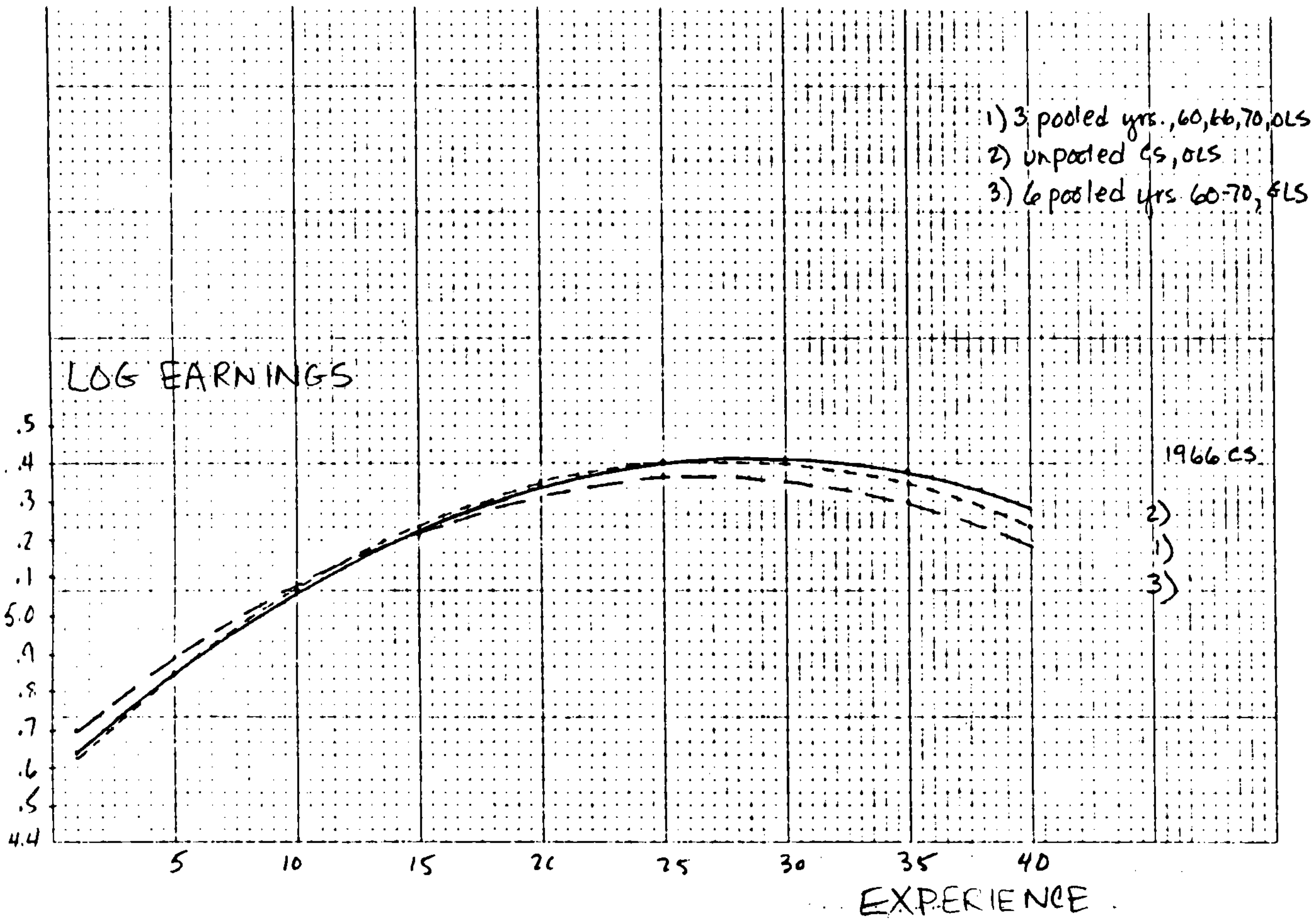
INDEP. VAR.	DEPENDENT VARIABLE IS LOG EARNINGS															
	GLS CHEMISTRY		GLS PSYCHOLOGY		GLS BIOLOGY		GLS PHYSICS		GLS EARTH SCI		GLS MATH		GLS ALL FIELDS		OLS ALL FIELDS	
	COEF	|t|	COEF	|t|	COEF	|t|	COEF	|t|	COEF	|t|	COEF	|t|	COEF	|t|	COEF	|t|
CONST	4.343		4.466		4.351		4.434		4.433		4.378		4.418		4.054	290.0
YR62	.0881	12.1	.0846	9.6	.0722	9.6	.0627	6.9	.0739	6.2	.0852	8.0	.0813	22.8		
YR64	.1590	16.5	.1700	12.5	.1509	13.1	.1312	10.1	.1859	10.8	.1573	9.3	.1621	31.4		
YR66	.2271	17.4	.2530	12.7	.2021	12.0	.1745	9.4	.2546	10.5	.2189	8.7	.2270	30.8	.209	28.8
YR68	.2767	16.1	.3360	12.3	.2625	11.3	.2083	8.3	.3214	9.8	.2575	7.4	.2865	28.7		
YR70	.2536	11.6	.3373	9.5	.2488	8.2	.1572	4.8	.2961	6.9	.2198	4.9	.2685	20.8	.245	42.9
PRE EXP=(P)	.0116	5.8	.0129	6.0	.0094	3.5	.0187	7.1	.0202	5.9	.0082	2.3	.0160	15.3	.0128	9.3
PxE	.0002	2.0	-.0002	1.6	-.00008	.6	-.0005	3.4	-.0004	2.1	-.0001	.8	-.0002	3.2	-.00038	3.7
EXP=(E)	.0533	33.4	.0519	20.8	.0631	28.9	.0613	26.3	.0478	15.0	.0589	18.1	.0539	57.9	.0644	47.7
BRK=	-.0095	1.5	-.0163	1.4	.0511	8.0	.0334	2.0	-.0179	.7	-.0124	.6	.0064	1.7	.0440	10.2
BRKxE	.0001	.3	-.0001	.2	-.0008	2.4	-.0019	3.0	-.0007	.7	.0004	.5	.0001	.7	-.0021	7.9
AGEHD	.0049	2.8	.0039	2.2	.0072	3.2	-.00007	.03	.0017	.5	.0056	1.7	.0012	1.3	.0088	7.5
AGEHDxE	-.0013	14.6	-.0011	11.3	-.0015	13.1	-.0011	8.0	-.0007	4.0	-.0012	6.8	-.0010	21.5	-.0013	14.6
E^2	-.0006	10.3	-.0008	7.5	-.0006	6.7	-.0004	4.6	-.0008	6.1	-.0004	3.0	-.0006	16.5	-.0008	11.2
TOPSC	-.0117	1.1	.0037	.2	-.0196	1.4	.0150	1.1	-.0249	1.3	.0033	.2	.0056	.9	.012	.9
TOPSCxE	.0018	3.6	.0013	1.7	-.0002	.3	.0016	2.3	.0015	1.5	.0010	1.0	.0012	4.1	.0015	1.7
YHDxE	.00033	3.2	.000005	.03	.00055	3.5	.00079	4.9	-.00015	.7	.00076	3.3	.00035	5.4	.00027	3.7
FEMALE	-.1590	4.1	-.0956	3.7	-.1983	6.3	0	0	0	0	0	0	-.1853	10.8	-.124	5.3
FEMxE	-.0015	.8	-.0048	3.6	-.0013	.8	0	0	0	0	0	0	-.0012	1.5	-.006	3.9
PRIV=(PI)	.3662	26.7	.3198	4.3	.3294	10.3	.4122	19.2	.2614	7.2	.5785	11.9	.3326	39.6	.393	26.0
GOVT=(G)	.2314	8.4	.0347	1.5	.0259	1.0	.1886	5.9	.0903	3.1	.2976	3.8	.0912	7.9	.068	4.2
UNSTAB=(U)	.2298	14.0	.0234	1.2	.0232	1.3	.2837	15.9	.1512	5.0	.2003	6.9	.1566	19.2		
PIxE	-.0016	1.6	.0033	.5	-.0082	3.2	-.0018	.9	.0030	.9	-.0095	1.8	-.0038	5.9	-.006	7.1
GxE	-.0051	2.8	-.0024	1.0	.0020	1.1	-.0021	.9	-.0032	1.6	.0006	.1	-.0006	.6	.0022	2.5
UxE	.0001	.1	-.0004	.2	.0022	1.5	-.0034	2.3	-.0022	.7	-.0027	1.1	-.0007	1.0		
PIx62	-.0384	4.9	.1072	2.5	.0101	.6	-.0088	.7	-.0463	2.1	.0054	.2	-.0249	5.2		
PIx64	-.0742	8.8	-.0381	.8	-.0150	.7	-.0512	3.4	-.1395	5.6	.0003	.01	-.0676	12.9		
PIx66	-.1296	13.5	-.1244	2.1	-.0316	1.4	-.0982	5.7	-.1983	7.0	-.0597	1.5	-.1169	19.6	-.097	6.2
PIx68	-.1729	15.8	-.1276	1.8	-.0679	2.6	-.1510	7.5	-.2359	7.2	-.0943	1.9	-.1668	24.3		
PIx70	-.1654	13.3	-.1587	1.9	-.0515	1.7	-.1528	6.5	-.2188	5.8	-.0917	1.6	-.1636	20.8	-.110	7.1
Gx62	-.0407	2.8	-.0232	1.7	-.0231	1.7	-.0400	2.0	-.0324	1.8	-.0501	1.3	-.0350	5.2		
Gx64	-.0045	.3	-.0191	1.2	.0180	1.2	-.0075	.3	-.0307	1.6	-.0122	.3	-.0123	1.7		
Gx66	-.0147	.8	-.0582	3.0	-.0001	.01	-.0030	.1	-.0326	1.5	-.0303	.6	-.0333	4.0	-.026	1.5
Gx68	-.0469	2.3	-.0629	2.8	-.0369	1.9	-.0389	1.4	-.0483	2.0	-.0707	1.3	-.0593	6.3		
Gx70	.0314	1.4	-.0326	1.2	.0175	.8	.0244	.8	.0463	1.7	-.0333	.5	-.0038	.4	-.001	.1
Ux62	-.0291	3.0	.0114	1.0	.0261	2.4	.0048	.4	-.0154	.8	.0389	2.3	.0013	.3		
Ux64	-.0736	6.8	.0038	.3	.0260	2.1	-.0318	2.6	-.0552	2.5	.0170	.9	-.0249	4.5		
Ux66	-.1323	10.9	-.0080	.5	.0247	1.8	-.0844	6.0	-.0714	2.7	-.0077	.4	-.0573	9.2		
Ux68	-.1657	12.0	-.0098	.6	.0076	.5	-.1288	8.0	-.0980	3.2	-.0583	2.3	-.0861	12.0		
Ux70	-.1805	11.4	-.0028	.1	.0111	.6	-.1301	7.0	-.0641	1.8	-.0656	2.2	-.0896	10.8		
R^2	.533*		.671*		.668*		.624*		.676*		.701*		.608*		.634	

*Equations estimated in deviation form with constant calculated separately. R^2 values are weighted but exclude the contribution of the constant.

Appendix Table 3: Regressions by Estimation Technique - Chemistry

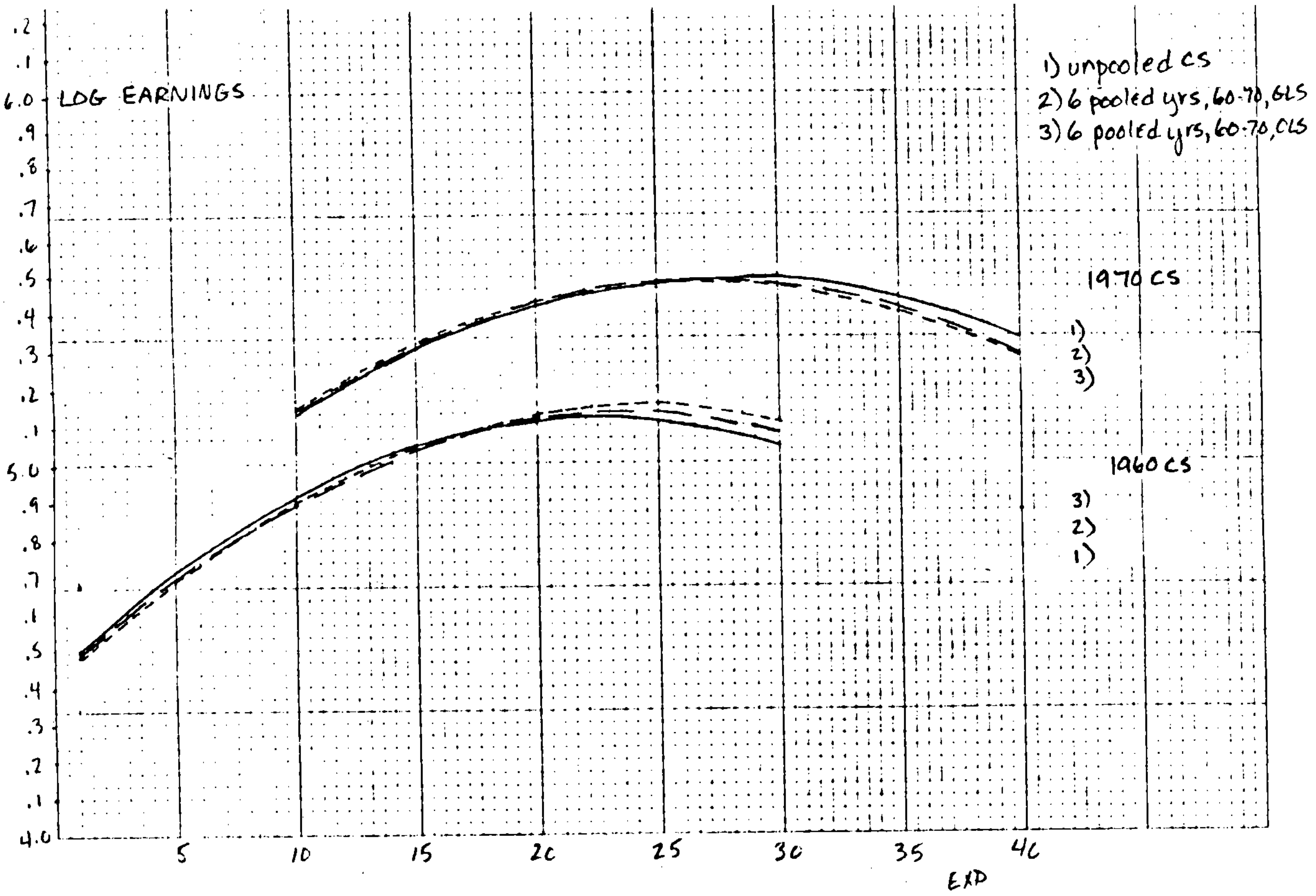
	OLS		GLS VARCOMP		GLS DURBIN	
	COEF	|t|	COEF	|t|	COEF	|t|
CONST	4.343	342.	4.343		4.346	223.
YR62	.0880	8.1	.0881	12.1	.0892	14.0
YR64	.1588	13.9	.1590	16.5	.1614	15.7
YR66	.2267	18.0	.2271	17.4	.2311	16.0
YR68	.2759	19.0	.2767	16.1	.2825	14.8
YR70	.2525	14.8	.2536	11.6	.2615	10.7
PRE EXP=P	.0157	10.4	.0116	5.8	.0128	5.0
P x EXP	-.0001	.9	.0002	2.0	.0001	.8
EXP=E	.0532	52.0	.0533	33.4	.0531	29.6
BRK	.0078	1.7	-.0095	1.5	-.0005	.1
BRK x E	-.0011	4.0	.0001	.3	-.0004	1.0
AGEHD	.0032	2.4	.0049	2.8	.0042	1.9
AGEHD x E	-.0012	14.3	-.0013	14.6	-.0013	9.5
E^2	-.0006	9.8	-.0006	10.3	-.0007	8.3
TOPSCH	-.0111	1.5	-.0117	1.1	-.0130	1.0
TOPSC x E	.0017	4.3	.0018	3.6	.0019	2.8
YHD x E	.00033	4.5	.00033	3.2	.00030	2.6
FEMALE	-.0987	3.7	-.1590	4.1	-.1527	3.2
FEM x E	-.0056	3.9	-.0015	.8	-.0018	.7
PRIVATE INDUSTRY = PI	.3684	36.0	.3662	27.7	.3599	25.4
GOVT = G	.2268	11.4	.2314	8.4	.2238	8.0
UNSTAB = U	.2309	18.3	.2298	14.0	.2258	13.3
PI x E	-.0018	3.6	-.0016	1.6	-.0010	1.0
G x E	-.0049	5.5	-.0051	2.8	-.0045	2.5
U x E	.0001	.1	.0001	.1	.0006	.5
PI x 62	-.0383	3.1	-.0384	4.9	-.0396	6.1
PI x 64	-.0742	5.9	-.0742	8.8	-.0768	8.4
PI x 66	-.1295	10.1	-.1296	13.5	-.1334	11.8
PI x 68	-.1728	13.2	-.1729	15.8	-.1780	13.5
PI x 70	-.1652	12.4	-.1654	13.3	-.1717	11.5
G x 62	-.0406	1.7	-.0407	2.8	-.0416	3.4
G x 64	-.0042	.2	-.0045	.3	-.0062	.4
G x 66	-.0144	.6	-.0147	.8	-.0173	.8
G x 68	-.0465	1.9	-.0469	2.3	-.0503	2.0
G x 70	.0319	1.3	.0314	1.4	.0271	1.0
U x 62	-.0293	1.8	-.0291	3.0	-.0302	3.7
U x 64	-.0739	4.6	-.0736	6.8	-.0758	6.5
U x 66	-.1327	8.2	-.1323	10.9	-.1356	9.5
U x 68	-.1663	10.0	-.1657	12.0	-.1701	10.2
U x 70	-.1812	10.7	-.1805	11.4	-.1859	9.8
R^2	.478		.533*		.957	

*Equations estimated in deviation form with constant calculated separately. R^2 values are weighted but exclude the contribution of the constant.



Appendix Figure 1. Comparison of Estimated Profiles by Data Source - 1966 Cross Sections.

All Fields - Academics



Appendix Figure 2. Comparison of Estimated Profiles by Estimation Techniques - Longitudinal Data

Physics - Academics