Volume Title: Higher Education and Earnings: College as an Investment and Screening Device

Volume Author/Editor: Paul J. Taubman, Terence Wales

Volume Publisher: NBER

Volume ISBN: 0-07-010121-3

Volume URL: http://www.nber.org/books/taub74-1

Publication Date: 1974

Chapter Title: Appendix I: Characteristics of the Residuals

Chapter Author: Paul J. Taubman, Terence Wales

Chapter URL: http://www.nber.org/chapters/c3664

Chapter pages in book: (p. 231 - 244)

Appendix I: Characteristics of the Residuals

In this appendix we examine some of the properties of the observed residuals (e_i) as an aid to evaluating our regression results. Any of our equations may be written in matrix form as $y = XB + u$, with the least-squares estimate of B denoted as b. Assuming that $E(b) = B$, then b is also the most efficient (linear unbiased) estimator of B, provided $E(uu') = \sigma^2 I$.[1] If, in addition, the individual elements u_i are normally distributed with a mean zero and variance σ^2, the least-squares estimates of B and σ are also the maximum-likelihood estimates, and the assumptions necessary to use the standard t and F tests (used in Chapter 5) are met. In the following analysis, we use the estimated errors to determine first if $E(uu') = \sigma^2 I$ and then if the u_i's are normally distributed.

In a general sense, heteroscedasticity exists whenever $E(uu') = \Omega \neq \sigma^2 I$. However, there is no reason to suppose in our case the $E(u_i u_j) \neq 0$, since we are considering unrelated individuals in a cross section. Therefore, we test the null hypothesis that $E(u_i^2) = \sigma^2$ for all i, against the alternative $E(u_i^2) = \sigma_i^2$, $\sigma_i^2 \neq \sigma_j^2$. Since we have repeated observations on X_i, that is, a number of people with the same X_i, we could develop an unbiased estimate of σ_i^2 from $e_i'e_i$. Of course, we would have to restrict ourselves to instances in which all the X_i's were the same for a group of people large enough to obtain reliable estimates of σ_i^2. We decided instead to use the entire sample in the following way. We divide the data into four education groups (with all graduate students combined) and five ability groups. Within each of the 20 possible cells, we compute $(\Sigma \bar{e}_i^2 - N_i \bar{e}_i^2)/(N_i - K)$

[1]The observed residuals cannot be used to test for bias because a property of least-squares residuals is that $(X'X)^{-1} X'e = 0$.

$= 1$, where K is the number of parameters in the equations and $\bar{e}$ is the mean error in the cell.[2] We also make this calculation for aggregated education cells, that is, for each education level after summing over all ability groups, and for aggregated ability cells.[3] In Tables I-1 and I-2 we present these estimates for 1955 and 1969.

All cells contain a large number of observations (see section A of both tables). In both 1955 and 1969, only two cells have less

[2]Estimates of $\bar{e}$ are obtained from equation 5 in Table 5-3 (pp. 82–85) and equation 5 in Table 5-7 (pp. 97–99).

[3]Of course, since the *b*s are estimated from the whole sample, the errors in the different cells are not independent; however, following the usual procedure as described in footnote 1 (of this chapter), we shall ignore such nonindependence.

TABLE I-1
Estimated variance by education and ability, 1955

	Ability: Y_{55}					
	Q_1	Q_2	Q_3	Q_4	Q_5	*Total*
A. Number						
High school	238	221	174	151	102	886
Some college	201	199	193	211	170	974
College degree	128	179	240	270	378	1,195
Some graduate	78	89	130	171	225	688
Total	640	588	737	803	875	3,743
B. Average Error = $\bar{e}$						
High school	15	1	18	−16	1	5
Some college	−13	33	−9	−17	21	2
College degree	18	−10	−6	14	10	5
Some graduate	−27	14	33	26	−19	5
Total	2	9	6	2	4	4
C. Variance (in thousands of dollars) = $(\Sigma e_i^2 - N_i\bar{e}_i^2)/N_i - K)$						
High school	53	53	65	52	51	51
Some college	68	141	71	65	207	98
College degree	57	51	49	109	84	70
Some graduate	43	68	78	74	65	60
Total	52	73	58	74	93	70

than 100 people.[4] The average errors vary widely across cells.[5] Although $\bar{e}$ must be zero (except for rounding) over the whole sample, there is no such restriction within each ability-education cell. However, if any $\bar{e}$ were significantly different from zero, then in our regression analysis we would have found an interaction between the education and ability levels corresponding to the cell.[6] None of the $\bar{e}$s is significantly different

[4] The reader is reminded that education changed between 1955 and 1969 and that there were different numbers of zero-income respondents who were dropped in the two years.

[5] Because of the sample size, we were not able to obtain residuals concurrent with the regression estimates. In the subsequent calculations of the residuals, we rounded all coefficients to one decimal, which created a small average rounding error of \$5.5 and \$−2.4 in 1955 and 1969, respectively.

[6] If the errors are normally and independently distributed, then the sum of T items would be distributed as $N(0,\sigma^2/T)$. A t test can be used to determine if any $\bar{e}$ is significantly different from zero.

Ability: log Y_{55}						*Ability: Y_{55} with Q variable*					
Q_1	Q_2	Q_3	Q_4	Q_5	*Total*	Q_1	Q_2	Q_3	Q_4	Q_5	*Total*
238	221	174	151	102	886	212	200	158	136	970	803
201	199	193	211	170	974	183	177	177	199	159	895
128	179	240	270	378	1,195	123	169	226	253	364	1,135
73	89	130	171	225	688	70	84	122	160	213	649
640	688	737	803	875	3,743	588	630	683	748	833	348
.009	−.010	.025	−.028	.011	.001	21	−5	23	−1	10	10
−.022	.010	−.009	−.020	.011	−.007	−8	24	5	6	22	9
.022	−.011	−.019	.009	.003	−.0004	18	1	−8	31	11	11
−.049	.016	.033	.034	−.024	.003	4	17	31	26	−15	10
−.005	−.001	.003	−.0004	−.002	−.001	10	8	9	17	6	10
.136	.134	.150	.138	.148	.127	36	36	44	53	32	36
.150	.288	.146	.155	.176	.167	51	84	44	40	17	68
.144	.110	.121	.145	.149	.127	42	47	41	89	61	55
.154	.151	.131	.130	.167	.130	43	47	59	64	53	48
.130	.159	.125	.132	.147	.136	38	48	41	59	70	52

TABLE I-2 ***Estimated variance by education and ability, 1969***

	Ability: Y_{69}					
	Q_1	Q_2	Q_3	Q_4	Q_5	Total
A. Number						
High school	219	214	162	152	92	839
Some college	202	179	208	208	162	959
College degree	124	160	216	263	352	1,115
Some graduate	98	121	161	195	285	860
Total	643	674	747	818	891	3,773
B. Average Error = $\bar{e}_i$						
High school	−47	12	39	15	−31	−2
Some college	65	11	−89	−24	39	−2
College degree	65	−6	66	−58	−24	−2
Some graduate	−125	43	−2	83	11	−2
Total	−2	−2	−2	−2	−2	−2
C. Variance (in ten thousands of dollars) = $(\Sigma e_i^2 - N_i\bar{e}_i^2)/(N_i - K)$						
High school	28	75	95	107	70	64
Some college	131	92	61	76	187	97
College degree	151	83	135	89	101	100
Some graduate	40	73	89	111	124	90
Total	78	73	88	87	112	87

from zero, although in 1969 they are almost significant in several of the graduate-ability cells.[7]

It is instructive to examine the general pattern of the variances. In 1955, the estimated variances of monthly earnings in the 20 cells range from a low of \$43,000 to a high of \$207,000, with only three estimates above \$100,000. In 1969, the low is \$280,000 and the high is \$1,868,000, with only four lying outside the range of \$1,000,000 ± \$300,000. A clearer picture of the relationship of the variance to ability and education is obtained by considering the "total" row and columns. In 1955, as ability

[7] The dummy variables in equations 13 and 14 in Table 5-7 indicate that the graduate–high-ability cells are significantly different from the graduate—low-ability and the high-ability—other education cells. This result merely confirms what we discussed in the text.

Ability: log Y_{69}					
Q_1	Q_2	Q_3	Q_4	Q_5	Total
219	214	162	152	92	839
202	179	208	208	162	959
124	160	216	263	352	1.115
98	121	161	195	285	860
643	674	747	818	891	3.773
−.069	−.006	−.039	−.002	.137	.003
−.036	−.012	−.023	−.012	.104	.0002
−.087	−.027	.018	−.063	.074	−.002
−.138	−.047	−.026	−.001	.076	−.002
−.073	−.020	−.002	−.024	−.087	.0003
.178	.229	.272	.279	.298	.218
.283	.257	.210	.239	.327	.239
.346	.209	.236	.228	.232	.226
.168*n*	.188	.206	.212	.208	.186
.218	.204	.211	.218	.229	.215

increases from Q_1 to Q_5, the variance increases from $52,000 to $93,000, although not monotonically. The highest variance in the education column is for the some-college group, while the lowest is in the highest education group. In 1969, variances increase with ability from $78,000 to $1,112,000, while with regard to education only the high school category is far from the average. In general, then, there does appear to be some relationship between the variance and education and ability.

We use a chi-square test developed by Bartlett (1937) to test the null hypothesis that the variances in all the cells are drawn from the same population.[8] The results of this test are given in

[8]The test statistic is $Z = (A \ln v - \Sigma a_i \ln v_i / C$ where $C = 1 + [\Sigma(1/a_i) - 1(A)/3\ (k - 1)]$; $A = a_i$; $v = a_i v_i / A$; v_i is the estimated variance in the ith cell; and a_i is the degrees of freedom in the ith cell. Z is distributed approximately as chi-square with $k - 1$ degrees of freedom.

TABLE I-2 (continued)

	Ability: Y_{69} with Q variable					
	Q_1	Q_2	Q_3	Q_4	Q_5	Total
A. Number						
High school	203	202	141	133	890	768
Some college	179	170	193	195	151	888
College degree	115	150	204	243	328	1,040
Some graduate	910	108	145	177	265	786
Total	588	630	683	748	833	3,482
B. Average Error $= \bar{e}_i$						
High school	−52	31	−19	93	60	0
Some college	58	−38	56	25	38	3
College degree	11	8	94	48	39	−2
Some graduate	−22	−27	−22	55	−13	−3
Total	−2	−4	3	20	−18	0
C. Variance (in ten thousands of dollars) $= (\Sigma e_i^2 - N_i\bar{e}_i^2/N_i^2/N_i - K)$						
High school	22	53	34	96	41	42
Some college	70	55	38	45	147	61
College degree	89	74	108	62	70	72
Some graduate	33	58	71	96	105	74
Total	46	53	59	65	85	62

Table I-3. In 1955, the test statistic for all cells is 252. For 19 degrees of freedom, the chi-square value that will be exceeded only 10 percent of the time (if all the variances are *drawn from the same population*) is 27.2.[9] Thus, we reject the null hypothesis of homoscedasticity for all education and ability cells. Further, performing this test for the education cells (after summing over ability) or ability cells (after summing over education) we still reject the hypothesis of homoscedasticity. Indeed, in 1955 (using equation 5 in Table I-3) we accept the null hypothesis only for the Q_1 column entries and the high school and graduate row entries.

In 1969, the test statistic over all cells is 244, which also exceeds the chi-square value at the 10 percent level. We reject

[9]The 5 percent level is 30.1.

the null hypothesis for the education cells and ability cells (after summing over ability and education respectively), although the test statistics are smaller than in 1955. The only instance in which we would accept the null hypothesis is for the variances in the Q_2 column in Table I-2.

Since the equations on which most of our analysis is based do not meet the necessary conditions for our estimates to be most efficient, it is necessary to consider whether alternative, more efficient estimates can be developed or, in other words, whether the heteroscedasticity has important implications for our results.

One common way to eliminate heteroscedasticity is to assume that the proper specifications of the equations is $\ln Y = \delta \ln X + v$, where v is normally distributed. The variances by education and ability for log equations are also given in Tables I-1 and I-2 and are tested for homoscedasticity in Table I-3.[10] In the log equations the results in Table I-3 are more favorable in both 1955 and 1969 to the null hypothesis than in the earlier equations, in that nearly all the test statistics are smaller. Even with these equations, however, we reject the null hypothesis over all cells, since the statistics of 46 and 67 exceed 27.2. In 1955 and 1969, we also reject the null hypothesis when testing the education cells, but we do not reject the null hypothesis when testing the corresponding ability cells in 1969. For the individual ability and education columns and rows, we reject the null hypothesis at the 10 percent level five out of nine times in 1969 and two out of nine times in 1955. Thus, while the log equations improve matters, the variances are still heteroscedastic.

Is it worthwhile, then, to analyze in detail the log equations, which are somewhat better in an efficiency sense than the ones in the text?[11] The following reasoning suggests that such a substitution is not worthwhile. The log equations yield estimates of the difference in log Y arising from education, ability, and so on. When Y varies over individuals, the average of the sum of the changes in the log of Y is not equal to the log of the differences in average income at the two education levels. Thus,

[10] Actually, we took only the log of Y. Since nearly all the other variables are zero-one dummies, logs of the independent variables are not necessary.

[11] Since nearly all our variables are entered in dummy-variable form and since we have tested for interactions, our equations with Y are as nonlinear as those with log Y.

TABLE I-3 *Test of equal variance in ability-education cells in 1955 and 1969*

Groups tested	Y_{55}		*ln* Y_{55}		Y_{55} *with Q variable*	
	Degrees of freedom	*Test statistic*	*Degrees of freedom*	*Test statistic*	*Degrees of freedom*	*Test statistic*
(1) All cells	19	252	19	67	19	249
(2) Education over all ability	3	108	3	27	3	84
(3) Ability over all education	4	74	4	14	4	88
(4) Education in Q_1	3	5*	3	1*	3	5*
(5) Education in Q_2	3	66	3	48	3	32
(6) Education in Q_3	3	10	3	3*	3	5*
(7) Education in Q_4	3	81	3	1*	3	33
(8) Education in Q_5	3	108	3	2*	3	101
(9) Ability in high school	4	3*	4	1*	4	10
(10) Ability in some college	4	97	4	31	4	121
(11) Ability in undergraduate	4	54	4	7*	4	43
(12) Ability in graduate	4	6*	4	4*	4	4*

*The null hypothesis is not rejected at the 10 percent level.

NOTE: At the 10 percent significance level, the value of the chi-square is 6.3, 7.8, and 27.2 for 3, 4, and 19 degrees of freedom.

we would have to convert the log of the geometric income differences to the arithmetic income differences. This conversion can be done in two ways. First, for every individual we could (1) add on to the log of his actual income the difference in log Y arising from education, (2) take the antilog of the new income, (3) find the difference in income, and (4) average over all individuals in a given beginning education level. Although this method would yield the correct answer for this sample, it need not be suitable for generalization to the census and other samples with different income distributions. Alternatively, if e^{v} is distributed log normally, it can be shown that $AM = GMexp(-\sigma^2/2)$, where AM and GM are the arithmetic and geometric means of income respectively. Unfortunately, as shown below, the distribution of e^{v} is not log normal, and we only have es-

Y_{69}		$ln\ Y_{69}$		Y_{69} *with Q variable*	
Degrees of freedom	*Test statistic*	*Degrees of freedom*	*Test statistic*	*Degrees of freedom*	*Test statistic*
19	244	19	46	19	296
3	29	3	15	3	72
4	44	4	3*	4	76
3	141	3	23	3	81
3	2*	3	3*	3	5*
3	39	3	4*	3	70
3	8	3	11	3	31
3	49	3	15	3	47
4	81	4	13	4	85
4	65	4	9	4	92
4	36	4	2*	4	19
4	44	4	3*	4	37

timates of σ_v^2; hence, the variances in our log equations would not give valid estimates of the variance associated with the arithmetic mean.

The results on heteroscedasticity based on equation 5, Table 5-3 (pp. 82–85), and equation 5, Table 5-7 (pp. 97–99), need not hold once we introduce the *Q* variable (the individual's residual from the other cross section). That is, suppose that there is an unobservable variable *P* that has a common mean but different variance in each ability-education cell. Assuming that *P* is a determinant of income in each year, our *Q* variable will eliminate at least part of its effect and the remaining error could be distributed homoscedastically. Unfortunately, as indicated in Table I-3, we still reject the null hypothesis of homoscedasticity except for four instances in 1969 and 1955.

All these tests suggest the existence of heteroscedasticity, and indeed we can explain to some extent why it is found in our sample. In Chapter 8 we presented estimates of the variance of the *e*s by occupation and education level. These variances differ by occupation and, to some extent, by education. But education and occupation, as well as IQ and occupation, are correlated. This suggests that we could reduce heteroscedasticity by including occupational dummies, but the inclusion of such variables will yield education coefficients inappropriate for the (direct) determination of the return to education.[12] Now let us turn to the implications of heteroscedasticity. As noted earlier, the existence of heteroscedasticity means that our estimating technique is inefficient. Since we are using regression analysis to accomplish a form of variance analysis, however, the inefficiency aspect is not as severe as usual. That is, in variance analysis, we are interested in the means of items in various cells. Suppose we only had two education classes; then even if the variance in the two classes were different, we would calculate the mean income in each cell as $(1/N)\ \Sigma Y_i$. Our regression analysis with dummy variables makes exactly the same calculation for mean income or differences in means except, of course, that we eliminate the effects of ability (and other) variables. Since these variables are not orthogonal to the education variables and since the effects of ability need not be the same in each education level, our estimates of mean income need not be efficient.

We did experiment with a generalized least-squares estimate to eliminate heteroscedasticity. For equation 5 in Tables 5-3 and 5-7, we weighted each observation by the reciprocal of the standard error of the ability-education cell in which the observation falls. As reported in Chapter 5, the 1955 coefficients are about the same, while their standard errors are smaller. In 1969, some of the coefficients changed slightly, but our basic conclusions remained unaltered.

The second general question to consider is whether the distribution of the errors is normal. To examine this question, we arrayed the errors monotonically and tabulated the number of

[12]See Chapter 2 for a discussion of this proposition.

residuals that fell within successive intervals of length $\sigma/2$.[13] The results for various equations are presented in Table I-4 for 1969 for the various ability and education cells. In both 1969 and 1955 the log equation has a median and mode less than zero and a large right-hand tail. In addition, the equation without Q has its median and mode less than zero. The right-hand tails reflect the fact that none of our equations will predict the earnings of those whose income is over $40,000. We also studied the distribution by ability and by education of those individuals whose residual exceeded 3.5σ. Generally, the educational and ability distribution of those individuals is about the same as in the sample; hence, being very successful is not a function of education or mental ability. Moreover, as shown in Table I-4, the tail and skewness can be found in each ability and education cell.

Finally, the information in Tables I-1 and I-2 can be used, albeit in a nonrigorous fashion, to discuss one other problem. There is some evidence in the literature that the income distribution (above some minimum income level) follows a Pareto distribution in which the *expected* value of the variance of income is infinite.[14] Even if the distribution of income is Pareto, the distribution of the error term, u, could be normal. However, if the distribution of the error term is also Pareto, then ordinary least squares is not an efficient estimating technique.

We believe that the above evidence strongly suggests that the error term is not distributed as Pareto. That is, if we took random drawings of the us and computed σ^2, we should not find the estimates converging to a single value as we increased our sample size, nor should we find the σ^2s of a given set of drawings following a particular pattern. As we increase the sample sizes by summing over the rows and columns in Tables I-1 and I-2, however, the estimates do converge. Moreover, the differences that remain follow the same pattern in 1955 and 1969 and are explainable by the occupational variations.

[13] We calculated the percentages in intervals of $0 \pm (k/2\sigma)$ and $^1/_4 \pm (k\sigma/2)$, where $k = 0, 1, \ldots$.

[14] Of course, in any finite sample, the formula for a variance could be used to obtain finite value.

TABLE I-4 Distribution of errors for Y_{55} and Y_{69}

Number in cell	Ph.D.	Master's	Some graduate work	B.A.	Some college	No college	Q_5	Q_4	Q_3	Q_2	Q_1
log Y_{55}	293	360	206	1,115	959	839	891	818	747	674	643
Range of σ											
$-2\frac{1}{2}$ to -2	.003	.000	.000	.000	.000	.000	.001	.000	.000	.000	.000
-2 to $-1\frac{1}{2}$	.017	.000	.000	.002	.000	.000	.004	.001	.001	.001	.000
$-1\frac{1}{2}$ to -1	.102	.027	.028	.047	.025	.004	.081	.026	.020	.012	.014
-1 to $-\frac{1}{2}$	.268	.159	.269	.280	.303	.177	.245	.282	.273	.230	.217
$-\frac{1}{2}$ to 0	.230	.387	.368	.342	.332	.505	.319	.363	.365	.405	.433
0 to $\frac{1}{2}$	.139	.258	.170	.144	.168	.176	.149	.151	.174	.199	.186
$\frac{1}{2}$ to 1	.088	.056	.090	.072	.073	.058	.082	.065	.068	.062	.074
1 to $1\frac{1}{2}$	.078	.016	.042	.044	.033	.032	.050	.034	.037	.044	.022
$1\frac{1}{2}$ to 2	.017	.021	.009	.025	.023	.019	.026	.027	.021	.015	.016
2 to $2\frac{1}{2}$	.030	.005	.014	.011	.008	.009	.009	.018	.009	.006	.012
$2\frac{1}{2}$ to 3	.003	.005	.009	.014	.011	.007	.010	.016	.008	.007	.006
3 to $3\frac{1}{2}$	.007	.003	.000	.010	.009	.008	.007	.008	.011	.007	.006
$3\frac{1}{2}$ to 4	.003	.000	.000	.002	.000	.000	.006	.000	.000	.001	.003
4 to $4\frac{1}{2}$	.003	.000	.000	.004	.002	.001	.002	.000	.003	.001	.000
$4\frac{1}{2}$ to 5	.003	.005	.005	.000	.002	.000	.001	.001	.001	.003	.000
5 to $5\frac{1}{2}$	.000	.000	.000	.002	.001	.001	.000	.001	.001	.000	.002
$5\frac{1}{2}$ to 6	.000	.000	.000	.002	.001	.000	.000	.002	.000	.001	.000
6 to $6\frac{1}{2}$	.000	.000	.000	.001	.000	.000	.001	.000	.000	.000	.000
$6\frac{1}{2}$ to 7	.007	.003	.000	.001	.002	.000	.003	.001	.001	.000	.002
7 to $7\frac{1}{2}$	.000	.000	.009	.003	.004	.001	.003	.000	004	003	005
$7\frac{1}{2}$ to 8	000	000	000	000	000	002	000	001	000	000	000

log Y_{89}

Range of σ											
-4 to $-3\frac{1}{2}$	.000	.000	.000	.001	.000	.000	.001	.000	.000	.000	.000
$-3\frac{1}{2}$ to -3	.003	.000	.000	.001	.001	.000	.001	.000	.001	.000	.002
-3 to $-2\frac{1}{2}$	.007	.003	.005	.003	.000	.005	.001	.006	.000	.001	.000
$-2\frac{1}{2}$ to -2	.010	.006	.005	.013	.008	.002	.014	.008	.001	.006	.009
-2 to $-1\frac{1}{2}$	.050	.014	.028	.030	.030	.017	.042	.021	.029	.028	.012
$-1\frac{1}{2}$ to -1	.060	.066	.056	.085	.113	.085	.081	.090	.088	.082	.093
-1 to $-\frac{1}{2}$	.147	.141	.225	.175	.198	.147	.155	.191	.209	.188	.174
$-\frac{1}{2}$ to 0	.260	.302	.263	.257	.212	.291	.258	.255	.239	.261	.263
0 to $\frac{1}{2}$	.197	.251	.164	.176	.166	.175	.182	.172	.179	.188	.185
$\frac{1}{2}$ to 1	.130	.144	.141	.103	.115	.086	.116	.105	.107	.113	.110
1 to $1\frac{1}{2}$	.097	.046	.052	.073	.075	.058	.077	.066	.065	.055	.078
$1\frac{1}{2}$ to 2	.030	.023	.033	.038	.036	.039	.032	.038	.040	.044	.022
2 to $2\frac{1}{2}$	.013	.006	.014	.024	.020	.019	.020	.027	.013	.010	.022
$2\frac{1}{2}$ to 3	.007	.006	.005	.011	.015	.015	.009	.011	.012	.012	.016
3 to $3\frac{1}{2}$	.003	.003	.000	.004	.003	.006	.003	.004	.007	.006	.000
$3\frac{1}{2}$ to 4	.003	.003	.005	.004	.006	.001	.007	.002	.003	.003	.005
4 to $4\frac{1}{2}$	.000	.000	.005	.000	.002	.002	.000	.000	.003	.001	.003
$4\frac{1}{2}$ to 5	.000	.000	.000	.000	.000	.001	.000	.001	.000	.000	.000