

Education and Income 1993: Stream of Evidence

One of the truisms of modern life is that higher education is associated with higher economic return to the individual. It is, of course, a complicated relation: Even the averages will be affected by age, sex, occupation, parents' backgrounds and other variables. And within each group defined by age, sex, occupation, and so forth, there will be considerable variation around the average. But the general nature of the relation, more education more money, is so widely understood and, presumably, so strong that it is worth beginning with the unadorned data: Education by Income. These data are from the 1993 General Social Survey from the National Opinion Research Center. The data for 1607 adults indicate education and income (plus about 400 other indicators), where "education" is defined as years of school completed, and the income indicator used here is the respondent's personal income (not household income or wealth). The responses for education range from zero years of education to 20 while income is presented in twenty-one categories ranging from zero income to \$75,000 per year-plus.

Here are the first 10 rows of my spread sheet. It demonstrates some of the idiosyncracies of the culture of data analysis. (Data analysts, like another other profession have a culture, and like any culture it develops lags, which are holdovers from older technologies and just plain unthought-out practices that have become customary.)

Note, for example, the names of the variables "EDUC" and "RINCOM91". Not very long ago the costs of most things involved with computers were so high that it was the practice to cut corners. Little things, like exclusive use of upper case letters, saved money. Conventionally names were cut to a maximum of 8 characters, hence "EDUC", rather than the English "Education", and "RINCOM91", which is the Respondent's Income in 1991. Add to that the advantage of giving variables the same name this year as you they were given last year and

the year before, and the result is the chopped English commonly used for the names of variables.

Item	EDUC	RINCOM91	Lookup Income	Ln of Income (base e)
1	16	18	\$45,000	10.71
2	12	14	\$23,750	10.08
3	12	16	\$32,500	10.39
4	14	21	\$100,000	11.51
5	14	21	\$100,000	11.51
6	15	17	\$37,500	10.53
7	15	13	\$21,250	9.96
8	12	13	\$21,250	9.96
9	17	3	\$3,500	8.16
10	12	15	\$27,500	10.22

For the same reasons of practical necessity, necessity in an earlier era of computing, you didn't write \$45,000. That would take seven characters, including the comma. Instead, you wrote "18". The meaning of "18" offers nothing to the intuition of the data analyst, but "18" as compared to "\$45,000" saves five characters. So it was the practice to write "18" in the data set and create a look up table (in a "codebook") that decoded the symbols into data. Here for example, the lookup table was:

0 M NAP	13 \$20000-22499
1 LT \$1000	14 \$22500-24999
2 \$1000-2999	15 \$25000-29999
3 \$3000-3999	16 \$30000-34999
4 \$4000-4999	17 \$35000-39999
5 \$5000-5999	18 \$40000-49999
6 \$6000-6999	19 \$50000-59999
7 \$7000-7999	20 \$60000-74999
8 \$8000-9999	21 \$75000+
9 \$10000-12499	22 REFUSED
10 \$12500-14999	98 M DK
11 \$15000-17499	99 M NA
12 \$17500-19999	

Note the intentional lack of correspondence between the symbol “0” and an income of zero or between the symbol “0” and the lowest category of income. This custom dates back to a time when the absence of data might leave no mark on the data sheet. It is absolutely routine for respondents to refuse one or more questions (and, for that matter, to refuse the whole questionnaire). So you have to be prepared to distinguish between no datum and no income and you have to be prepared to catch mistakes when the distinction breaks down. So, it seemed wise to minimize errors by not using 0’s as data. Combine this confusion with an equally careful practice of intentionally attaching nonsense symbols to nonsense data and you have a mess. (Introducing “99” for No Answer means that if, somehow, that symbol “99” were used as data it would stand out in a stem and leaf or in a two-variable graph. The data analyst will see the stem and leaf, or the graph, know that something is very wrong, and quickly track down the error.)

There is no obvious convention for these symbols, so each study requires you to “look up” the symbols in the codebook and find out what they stand for. Here the symbol “0” stands for “M NAP”, that is, “Missing - Not Applicable.” Here “22” stands for “Refused”, “98” stands for “Missing - Don’t Know”, and “99” stands for “Missing - No Answer.”

You want all of this detail about “non-responses” for three reasons. First, you want to know how many people responded. The question may have been asked of 1,500 people. That does not mean that 1,500 people responded. Second, non-responses are very likely to be associated with particular values (usually extreme values) which would mean, in this case, that very low incomes and very high incomes are less likely to be recorded). And third, the reasons for non response are likely to be associated with other variables. So you have to be alert to the possible divergence between *observed* correlations (using the data) and true but unknown correlations that you *would* have seen if everyone had responded.

You also see the custom of placing numbers like income in categories. There are arguments for and against the use of categories

rather than raw income numbers. But whatever the argument, the use of categories, rather than income numbers, has consequences. In this case it means that all people with incomes above or equal to \$75,000 dollars are the same for purposes of analysis: \$75,000, \$100,000, \$1,000,000 — all equal. We can argue whether or not that is an appropriate decision (a debate sometimes described as the “One person, one vote” versus the “One dollar, one vote” debate), but with these data the decision has been made, the detail is gone, and it can not be restored.

Now, on to the analysis or, at least to the preparations for the analysis. Using my spread sheet program’s “lookup” function I have re-expressed income with approximate values appropriate to each category. I have substituted a nonsensical -999999 for all forms of missing data, and then removed all missing data to the bottom of the spread sheet by sorting the data. For income that removes 539 of the 1607 respondents, about one-third missing. For education that removes 8 respondents.

These deletions are certainly disconcerting and hard to check without analyzing the attributes of these people with respect to other variables (Are they disproportionately men or women? Are they disproportionately young, old, or in some other age group? Are they simply unemployed or unemployed outside the household?) The briefest check, within these data shows that the 539 missing income population has a mean education of 12.0 years of education as compared to 13.1 years for the whole 1607 population. Their standard deviation for education is 3.1 as compared to 3.0. That tells me that the histograms of the two distributions could look very similar, with the distribution of incomes offset to lower education. And although the difference is small, it is probably real, not the luck of the draw.¹

¹ I find that I am resorting to statistical knowledge that I have not included in the text. So, let me explain, or at least start to explain the difference between the standard deviation of the values of education and the standard deviation of *means* of the values of education. The mean education in this population is 13.1 years. That is a fact resulting from the obvious computation. The standard deviation of values around this mean is 3.0 years. Again, this is a fact resulting from the obvious computation.

For analysis of the data using both education and income, this reduces the number of adults from 1607 to 1017, removing 40% of the data. Presuming that these “missing” respondents are not a random sample of all respondents I am worried. But I have no effective way of

But consider: Suppose that I went out and collected new data for another 1607 people. I would get another mean and it would almost certainly be close to but slightly different from the mean I found in the first sample of 1607. If I did this again and again, getting a new sample and computing the mean in each sample I would get a lot of means, most of them close to the original but slightly different.

These means themselves would have a mean and a standard deviation. And if I were comparing one sample to another, asking whether one of these samples has a mean that is too much larger than the others, or too much smaller than the others to be ignored, then I would also need to know the standard deviation of the mean.

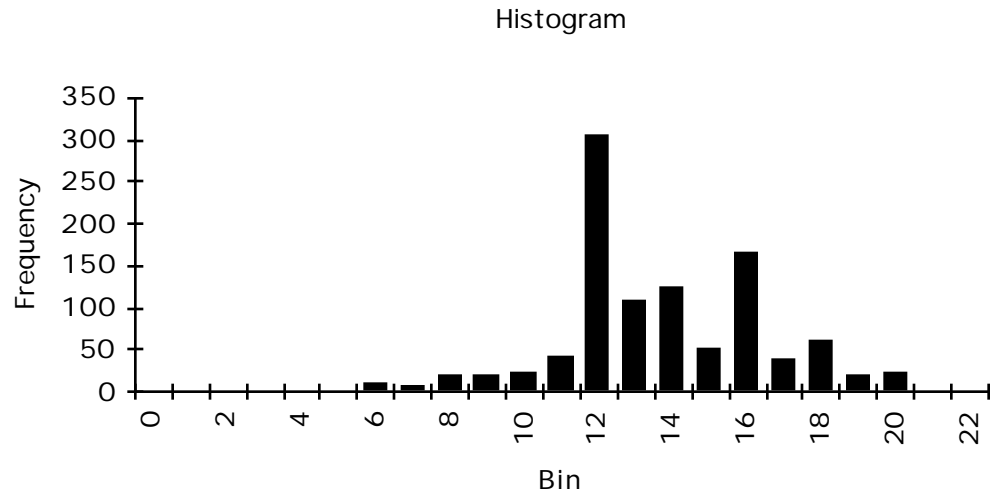
That is close to what we are doing here. We are comparing one group of 1607 with a mean of 13.1 years of education to a subset of 539 people who have a mean of 12.0 years of education. Are these two means close? I narrow the question by asking whether or not these two means are close as compared to the standard deviation *of the mean*.

Fortunately, statistics is able to estimate the standard deviation of the mean without the need to actually perform the experiment, without actually collecting new data and computing the mean, again and again and again. We know that the standard deviation of the mean is approximately the standard deviation of the values divided by the square root of the number of values. Here the standard deviation of the subset is 3.1. The number of values in the subset is 539. The square root of 539 is 23.2. So the standard deviation of the mean is $3.1/23.2 = .13$. That is my estimate for the standard deviation of the mean itself.

The difference between these two means is 1.1 year of education which is more than 8 times greater than the standard deviation of the mean. So, the difference in education between the subset and the whole is small, 1.1 year of education. But it is almost certainly a real difference, the difference is small but greater than I would expect just by the luck of the draw.

worrying about these numbers here. The correct way to check what is special about the missing people is to considering other variables available in the study — which is well beyond the scope of this exercise.)

The education variable (for the 1017 people) has decidedly non-bell-shaped exceptions:



These bumps are characteristic of education distributions, with bumps at the numbers of years that correspond to degrees. A bump at completion of grade school is no longer apparent in 1993. But there is a bump at 12, usually completing a high school degree in the U.S. and there is a bump again at 16, a college degree.

Is it symmetrical? That is a bit tricky because I don't really know how I want to qualify this question in order to accommodate these perfectly reasonable bumps. Let me take a look

Mid Values, Years of Education

Count, n=1015	From bottom	From Top	MidValue	
508	13	13	13	Median
254.5	12	16	14	Mid Quartile
127.5	12	17	14.5	Mid Eighth
64	10	18	14	Mid Sixteenth
32.5	7.5	19	13.25	Mid Thirty-Second

Mid Values, Square Root of Years of Education

Count, n=1015	From bottom	From Top	MidValue	
508	3.61	3.61	3.61	Median
254.5	3.46	4	3.73	Mid Quartile
127.5	3.46	4.12	3.79	Mid Eighth
64	3.16	4.24	3.57	Mid Sixteenth
32.5	2.73	4.36	3.54	Mid Thirty-Second

Mid Values, Natural Log of Years of Education

Count, n=1015	From bottom	From Top	MidValue	
508	2.56	2.56	2.56	Median
254.5	2.48	2.77	2.63	Mid Quartile
127.5	2.48	2.83	2.66	Mid Eighth
64	2.30	2.89	2.60	Mid Sixteenth
32.5	2.06	2.94	2.50	Mid Thirty-Second

There are enough values here to allow me to pursue quite a number of mid values without running out of data. The median, mid quartile and mid eighth do show a trend. But the mid sixteenth and mid thirty-second show the reverse.

I know that this is a bit of a mess, and I don't trust the top number — there are more than a few of us with greater than 20 years of education but we seem to have been left out or lumped in with our relatively uneducated friends with merely 20 years of education. So I think that the distribution has more of a tail than it is able to show with

these categories. Checking to see whether a re-expression would fix up the trend among the first three mid values, the curious result is that all three sets of mid values show the same thing. So the test for symmetry is beautifully indeterminate, quite willing to accept the original numbers, or the square roots, or the logs, or more. That is no help at all. I will start simply, using years of education.

The same procedure applied to the income distribution of this limited population yields

Mid Values, Respondent's Income

Count, n=1015	From bottom	From Top	MidValue	
508	\$21,250	\$21,250	\$21,250	Median
254.5	\$11,250	\$32,500	\$21,875	Mid Quartile
127.5	\$5,500	\$45,000	\$25,250	Mid Eighth
64	\$2,000	\$67,500	\$34,750	Mid Sixteenth
32.5	\$2,000	\$100,000	\$51,000	Mid Thirty-Second

Mid Values, Square Root of Respondent's Income

Count, n=1015	From bottom	From Top	MidValue	
508	145.77	145.77	145.77	Median
254.5	106.07	180.28	143.18	Mid Quartile
127.5	74.16	212.13	143.15	Mid Eighth
64	44.72	259.81	152.27	Mid Sixteenth
32.5	44.72	316.23	180.48	Mid Thirty-Second

Mid Values, Natural Log of Respondent's Income

Count, n=1015	From bottom	From Top	MidValue	
508	9.96	9.96	9.96	Median
254.5	9.33	10.39	9.86	Mid Quartile
127.5	8.61	10.71	9.66	Mid Eighth
64	8.16	11.12	9.74	Mid Sixteenth
32.5	7.60	11.51	9.56	Mid Thirty-Second

These numbers are strange: The distribution of incomes is not symmetrical. It has a tail toward the high values. That is reasonable. But the square roots of income lead to ambiguous indicators: The median, the mid quartile and the mid eighth drift slightly toward low values. The mid sixteenth and mid thirty-second move strongly toward the high values. Worse, the distribution of the logarithm of income indicates a tale to the left, if any. This is surprising: I have it in my head that income distributions will not be symmetrical measured in units of dollars but will be symmetrical measured in units of log dollars. That is what I expect. That is not what these data show. What's more, I am so sure of this that I question the data.

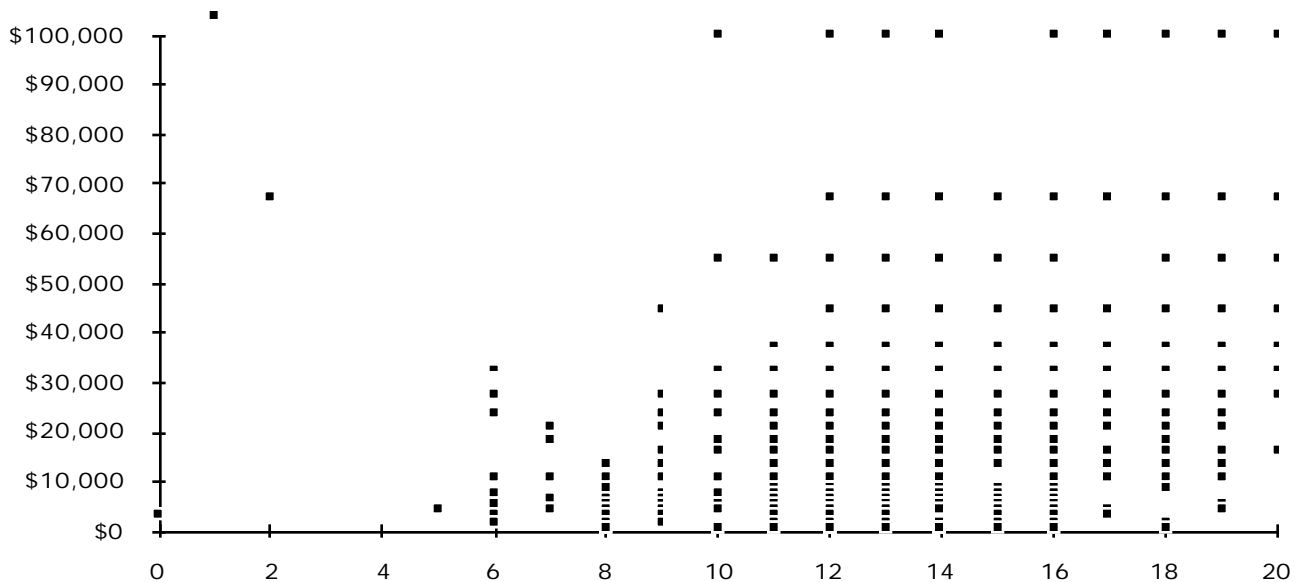
A close look at the data tells me that the unit of analysis in these data is the family, not the individual — it is a representative sample of family units from which the individual who is actually interviewed is chosen by ????. I don't know what. Either I am wrong about the shape of the U.S. income distribution or else there is something in these data that systematically under-represents high income individuals.

That tells me I am not going to reach a "final report" with these data: I will have to check both the income distributions and the educational distribution against other sources before I trust either distribution or their correlation. I'm also going to have to re-examine my own expectation about income distributions. When I think about income carefully, as I am now forced to do, I'm not sure what I should expect. After all, a very large part of the population has no cash income at all, and my intuition didn't cover that very realistic contingency. For the moment, if there is a bias I suspect that it is reducing the high income end of the distribution.

Proceeding rather tentatively, I am ready to look at the two variable distribution.

A graph of the relation between the two variables provides only a slight insight into the relation because of the grouping of the income data into categories. The grouping hides visual differences between the numbers of "dots" at each point on the graph. But the graph suffices to

bring home the fact that there is a wide range of incomes across the entire range of educational achievement, particularly among those who have completed high school, Figure 1.



Two obvious ideas would be expressed by “linear” relations in these data. One would be the idea that each additional year of education corresponds to a certain increase in average income, with the number of dollars per year appearing as the slope of the linear relation. The other idea would apply to a linear relation between education and the logarithm of income. In this case each additional year of education would correspond to a multiplication of the average income, a multiple that could also be expressed as a percentage.

While, in principle, it is not logically possible for both of these equations to be correct, the “noise” represented by the vertical scatter in either graph (either income by education or log income by education), is so great that it is not possible to choose between the two possibilities with these data.

I'm also still worrying about the well-behaved or not-well-behaved nature of these variables because it will affect the validity of using any linear technique. So I am going to pull out another property of well-behaved variables, namely that two well-behaved variables should be linearly related (if they are related at all). I'm going to use r to measure the strength of the linear correlation and see whether any combination of transformations has a useful effect on linearity (from which I will infer that I have found the well-behaved transformation).

Correlations	educ	sqrt educ	ln educ
inc	0.38	0.36	0.34
sqrt inc	0.38	0.37	0.35
ln inc	0.34	0.33	0.32

Squared Correlations	educ	sqrt educ	ln educ
inc	0.14	0.13	0.11
sqrt inc	0.15	0.14	0.12
ln inc	0.12	0.11	0.10

That tells me very little: Correlations with sqrt of income are slightly larger than others. Correlations with education are slightly larger than others. But in terms of "variance explained", the range is from 10 percent to 15 percent. And the most interesting fact is that all of them are low: Step away from the methodology: Using these data, at most "15 percent of the variation in income is predictable from education". This is an approximation to one of those facts of life that "everyone" knows to be true. Well, "approximately 15 percent of the variation in income is predictable from education"

Getting ready for an interim report, I will collect various means, and standard deviations, and then write what I can

	Income	Ln Income	Sqrt Income	Education	Ln Education	Sqrt Education
Mean	\$26,473	149.14	9.78	13.69	3.68	2.59

Standard Deviation	\$22,295	65.05	1.05	2.75	0.38	0.22
--------------------	----------	-------	------	------	------	------

So tentatively, I write what I can — definitely not ready for an external report. Definitely not ready to be the basis of any policy recommendations.

<p>The NORC General Social Survey for 1993 interviewed 1607 individuals selected from a national sample of households</p>	<p>Who, what, where, but also discretely noting my discomfort: I want to talk about individuals but I know that I can't. Yet I am not sure how to translate the findings for this sample, based on households, to a correct statement about individuals.</p>
<p>Respondents showed a mean education of __ years and a mean income of ____ .</p>	<p>Safe, but qualified because I said respondents. (I did not say that the mean education in the US was __ or that the mean income in the US was ____.</p>
<p>While there is a clear correlation between education and income, it is not strong. Estimating the regression line for income as a function of education, the line shows estimates that individuals with no formal education will have __ income while, on the average, each additional year of education corresponds to ____</p>	<p>Using the regression line</p>
<p>In these data, while there is a correlation for the population as a whole, it is not a relation that individuals can count on.</p>	<p>Trying to give some idea of the strength of the correlation</p>

<p>For example, the median income of individuals completing high school was \$18,750 in these data , indicating that 50% of these individuals have incomes at or exceeding this amount, 50% have incomes below \$18,750 By comparison, individuals with four more years of education, usually college graduates, have a median income of \$27,500 but 25 percent of these college graduates had incomes below the median for high school graduates, while 30 percent of the high school graduates had incomes exceeding the median for college graduates. That is, while higher education corresponds to higher income, on the average, there is no guarantee that a specific individual with higher education will have an income exceeding that of another individual with less education.</p>	<p>First try at an illustration. A pair of histograms, one for the income of high school graduates, one for the income of college graduates — both drawn to the same scale — should make this clear.</p>
<p>If we treat income and education together as indicators of social status, the high quartiles of these status variables pair off __ years of education with __ years of income, while the low quartiles of the status distribution pair off __ years of education with __ years of income, approximately __ dollars per year</p>	<p>Implicit use of the orthogonal least squares line. I am showing the correspondence between income and education without using words that imply that one variable depends on the other.</p>

<p>The indications from these data are suspect, requiring some verification before they are used as the basis for other work. First the income and education distributions need to be checked against US Census data, with a clear definition of just who's incomes we are talking about. (Are people who were unemmployed and had no other income excluded from the data or were they recorded as zero income.)</p> <p>The fact that this income distribution did not respond to a logarithmic transformation, by becomming bell shaped and symmetrical, leaves some suspicion of bias such that higher income individuals are less likely to be represented in these data. If so, it will have thrown off all the estimates of correspondence between income and education, probably reducingthe income numbers below their true values.</p>	<p>Clear doubts, plus specific reason for the doubts, and anticipation of specific consequences.</p>
--	--

<p>The best fit regression equation in dollars and years of education predicts about \$3,000 dollars per year of education completed, while the regression equation in log dollars predicts a 14% increase in income per year of education completed. The fact that these two numbers are not consistent points to the large variance of income in all educational groups, making it impossible to be more precise reasoning solely from these data. \$3,000 is a little under 14% of the average of \$26,500 For a non average education or income it is difficult to be more precise from these data. To give a bench mark to the size of these errors: If a forecaster were to predict an average income, predicting the same income for all persons, ignoring education, the standard deviation of the errors would be \$22,000. Using education to improve the prediction of income, the standard deviation of the regression errors would remain at about \$20,500.</p> <p>Nevertheless, the sense of the log equation is a more credible result. It predicts \$3,000 as the average income of persons with no education. By contrast, the equation in dollars predicts negative income, \$-14,742 for that same person.</p>	<p>Trying to get comfortable with the results.</p>
--	--

Macintosh HD:DA:DA IX:Volume II:320 Income by Education, March 19, 1997

Similarly, the implications of the error statistics also favor logarithms. The correlation of __ indicates that the variance of the residuals of income is __ percent of __ the variance of income. Where the variance of income is __, this implies that the standard deviation of the residuals would be __. Such a number gets us deeper into nonsense: Predicting a standard deviation of __ among persons with education of 0 and predicted income of __, a standard deviation of __ among persons with twelve years of education and predicted income of __, and the same standard deviation of __ among persons with twenty years of education and predicted income of __.

By contrast, using logs, the equation predicts a geometric

<p>In summary, on the average each year of education corresponds to about \$3,000 or 14% greater income. But the ranges of income at all educational levels requires a warning that there will be extremely large differences, both positive and negative among individuals with the same education and, therefore, between individuals with different educational backgrounds.</p>	<p>I would not want to leave the reader with the previous paragraph as a final comment.</p>

Re-work the spread sheet: First show the sorting of the data such that identical cases lie in adjacent rows. Then count these cases, displaying the result as a cross-tab. Then modify the spread sheet to compute mean squared error using frequencies. That should get the same job done as r and r -squared which I used in the previous draft. In this draft I can't use r -squared because I haven't prepared the background.

Modify after the work described above:

For the relation in dollars, the coefficient of linear correlation is .38 ; for the relation in log dollars, the coefficient of linear correlation is .34, “explaining” respectively, 14% or 12% of the variation in income.

Where the equation in dollars predicts \$3,013 dollars per year of education completed while the equation in log dollars predicts a 14% increase in income per year of education completed. While the log equation produces a more obviously sensible intercept, \$3,000 as the average income of persons with no education, it is not altogether obvious that the negative intercept for the dollar equation, \$-14,742, is altogether foolish, producing a “deficit” of \$15,000. Which should I use? The data are not strong enough to choose for me: I have a prejudice for using logs when I’ve got income data. that would lead me to the log equation. But, I also have a prejudice for sticking with common units if the data do not compel me to do otherwise. That would lead to dollars. But, then again, using dollars only starts out by looking like common sense, and then it presents me with the need to talk about negative income, which is possible but no longer consistent with my prejudice to keep it simple. So I will have to conclude that the data show about \$3,000 per year of education and about 14% per year of education. The fact that these two numbers are not consistent points to the large variance of income in all educational groups making it impossible to be more precise reasoning solely from these data. \$3,000 is a little under 14% of the average of \$26,500. For a non average education or income it is difficult to be more precise from these data. To give a bench mark to the size of these errors: If a forecaster were to predict an average income, predicting the same income for all persons regardless of education, the standard deviation of the errors would be \$22,000. Using education, the standard deviation of the regression errors would remain about \$20,500.