**EBM Application Assignment**

*Leandro Baldelli - Nishal Dave*

1. **Introduction**

This piece of work aims to provide exposure to the conditional quantile analysis carried out in the paper by Angrist et al. (2006) - (here referred to as ACFV).

ACFV is a theoretical paper that looks at properties of Quantile Regression under misspecification and, when trying to replicate it and expanding it, we are going to focus on its empirical application, especially the visual part related to Figure 2 from the paper.

Figure 2 from the ACFV empirical application section demonstrates how the returns to schooling coefficient changed during the last two decades of the past century, from 1980 to the year 2000.

Their original findings, as replicated below in *Figure A1*, show that through the 80s, returns to schooling significantly increased across the entire income distribution, with slightly higher gains at the top of the distribution in 1990. From the year 1990 to the year 2000, however, Quantile Regression shows that at the top of the income distribution returns to schooling had kept growing while at the bottom of it they had dropped.

Furthermore, from *Figure A2,* we can see that from 1990 and 2000, conditional wage inequality increased more in the upper half of the wage distribution than in the lower half, while between 1980 and 1990 the increase in inequality occurred in both tails.

We then decided to subdivide our work in expanding the ACFV paper into two main sections.

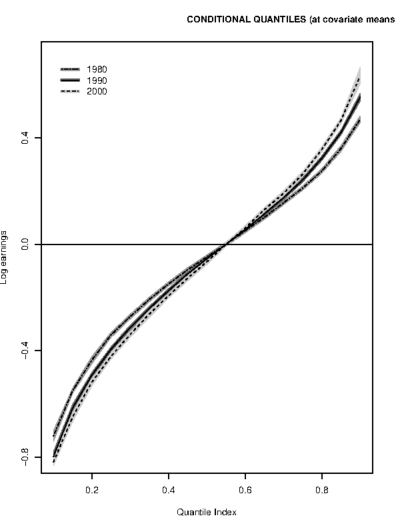
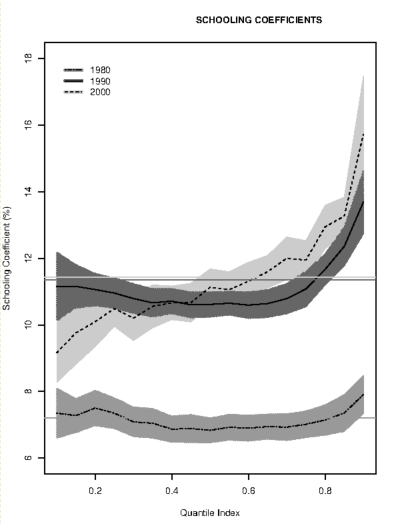
Firstly, we expanded the ACFV paper by collecting data from the 2010 and 2019 American Community Survey (ACS). Our objectives here were to see whether returns to schooling would have increased in the last two decades and whether wage inequalities would have changed.

To do this we collected data for the U.S.-born black and white men aged 40-49 with five or more years of education, with positive annual earnings, and positive hours worked in the year preceding the survey. We therefore used this data to extend the ACFV initial analysis to the years 2010 and 2019 and presented our results.

Additionally, we carried out a comparison between a sample of women and men, controlling for the same variables and characteristics as we did for the male sample. This was conducted to see whether there were any statistically significant differences in returns to schooling coefficients between the two samples in 2010 and 2019.

To carry out these analyses, however, we had to make sure that the data downloaded for the year 2010 and 2019 would match the same format as the datasets used for the original paper. This would allow us to use the original R codes for Figure 2 in the ACFV paper in the correct and unbiased way with the “new” data.

This process will be explained in detail in the next section.



***Figure A1 Figure A2***

In our conclusion we will summarise the results obtained from the new figures and consider any drawback and difficulties we had when carrying out this short analysis, as well as talking about any other possible further approach or expansion.

Finally, in the appendix, we will include the Python codes used for executing the operations of cleaning and formatting the data in a consistent way with the original datasets. The newly created datasets will be used in the original Figure 2 R script from the authors to create the new figures. The tweaked R script will not be included in this report because of its length and the fact that no major changes have been implemented from the original, apart from a small update of an obsolete code and changes in the graphs' visual outline.

1. **Data Cleaning and Formatting**

The data used for study this were collected from the IPUMS database for both 2010 and 2019, the additional gender variable has also been included in the analysis. For the analyses to be comparable the data were required to match the coding and format used by ACFV, this operation was carried out using Python along with Numpy and Pandas.

Firstly, it is important to ensure that the data in the dataset are accurate and free of imputations, such observations were labelled as “allocated” and subsequently dropped from the dataset to prevent any inconsistencies or outliers.

To remain comparative to ACFV, our study continues to use White and African-American men and women only. There were some limitations to using Hispanic individuals which are explained in section 5. The race variable has been filtered to take on one of the two ethnicities and then subsequently dropped after generating a new variable called “black” which takes on “1” if the individual is black and “0” if the individual is white.

The same principle as above has been applied to the gender variable, by way of creating a male dummy which takes on “1” if the individual is a male and “0” if they are a female.

As income is measured in nominal terms, for comparative purposes the yearly income was transformed into real terms in 1989 prices, by taking the ratios of the personal consumption expenditure index of 1989 and both 2010 and 2019 and multiplying these ratios by the yearly income this acts a deflationary measure and is useful in comparison.

Calculating weekly income had a challenging approach, as the data available was only in yearly wage for the year prior, to make a sound estimate further regarding the number of weeks worked per individual had been used. Taking an average of the weeks worked by an individual and using the ratio of yearly income to average weeks worked provided a more accurate estimate of an individual's weekly income. The average used varies slightly across the range of weeks, for the majority, the midpoint is a sensible estimate, except in the case of those working 50-52 weeks, where 52 weeks were used instead. This follows the logic that those in paid employment for 50-52 weeks are working for the whole 52 weeks, those working 50 or 51 weeks are assumed to account for a smaller proportion of the observed sample. The last step involves taking a natural log transformation of the weekly incomes of the observations.

Education data extracted from PUMS compared to ACFV had significantly different coding, which needed to be adjusted for, using a mapping procedure taken from the supplementary notes to ACFV. For many mappings there were no issues besides the mappings for associate and bachelor's degrees, both of which are represented by 2 values in ACFV, however the PUMS data does not distinguish these and only has one code for associate and bachelors' level of education each. To overcome this, a conservative approach has been taken to use 14 years of schooling for everyone with an associate's degree and 16 years of schooling for those with a bachelor's degree, as opposed to 15 and 17 years, which have been excluded from the mappings. With the education variable available, this allows easy calculation of the experience variable and its square.

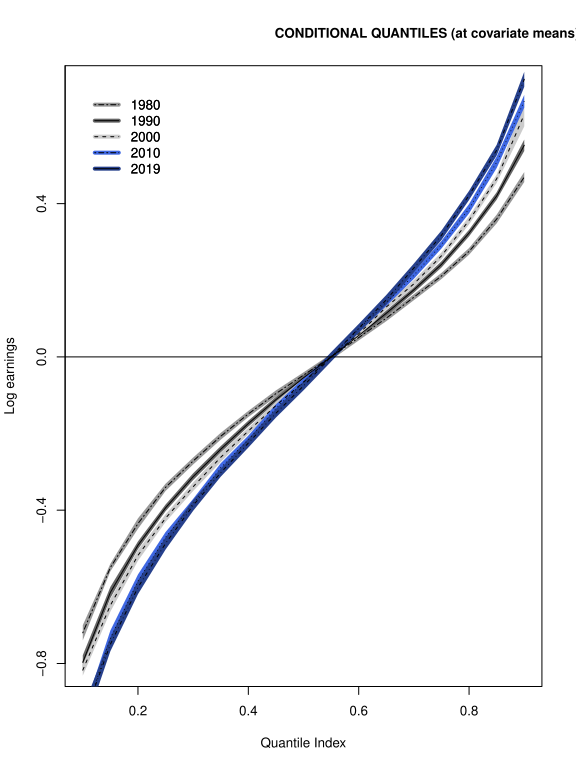
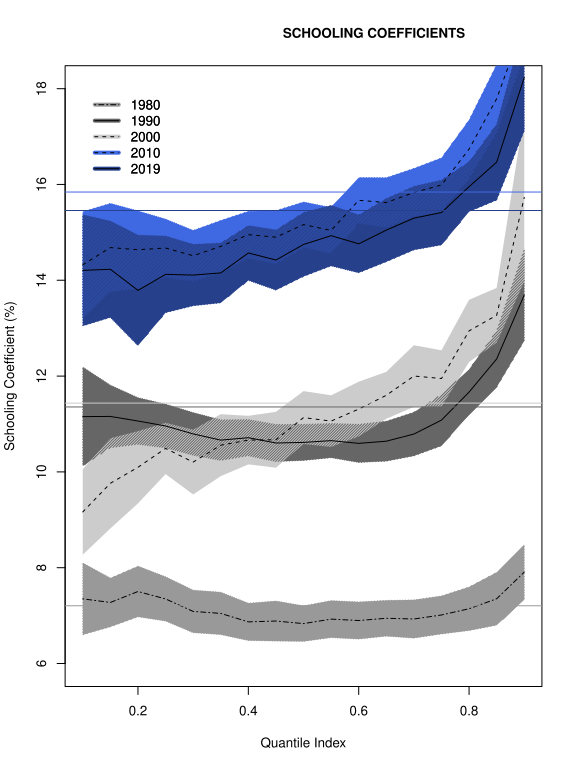
The extension of this study has allowed for the comparison of both men and women which differs from ACFV and therefore the data is finally split accordingly into four datasets to represent the year and gender of the individuals.

1. **ACFV 2010 and 2019 extension**

The *Figures B1* and *B2* below are an extension to ACFV and in addition to the original years observed, the updated plot shows the impacts of an extra year of schooling across the different quantiles for 2010 and 2019 for men. There is little change from 2010 to 2019, however compared to 2000 there is a significantly different impact of education. In addition the function has slightly flattened out implying that the impact of additional years of schooling is not as strong as it once was, this could be in line with recent progressions in career options available that facilitate higher earnings without the need for many years of education.

Towards the upper quantiles of earnings, education still plays a vital role, this is a characteristic that is likely to be quite resistant to time as many traditional jobs which pay large salaries are dominated by highly educated individuals, namely those in healthcare, law, engineering and finance.

Another interpretation of the conditional quantile function is the increasing steepness over the years shows that real income has worsened for the lower quantiles, whilst it has improved for the higher quantiles, this is evidence in favour of widening inequality over the last 40 years.



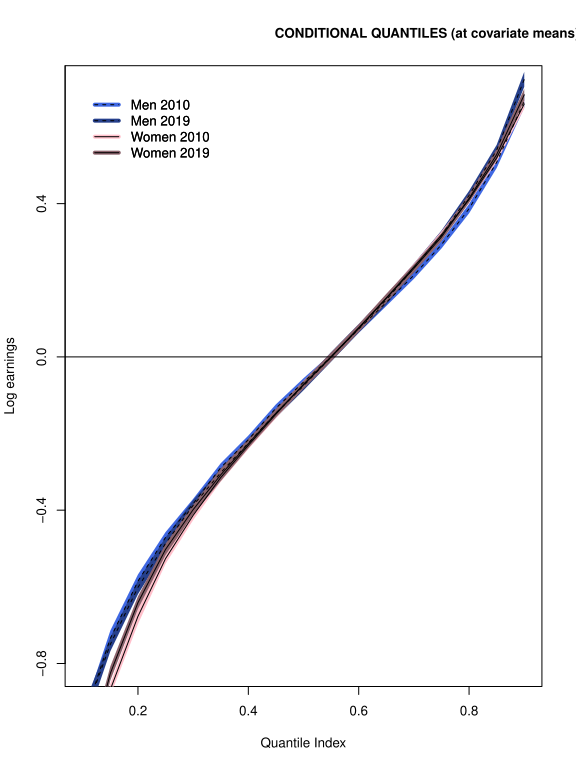
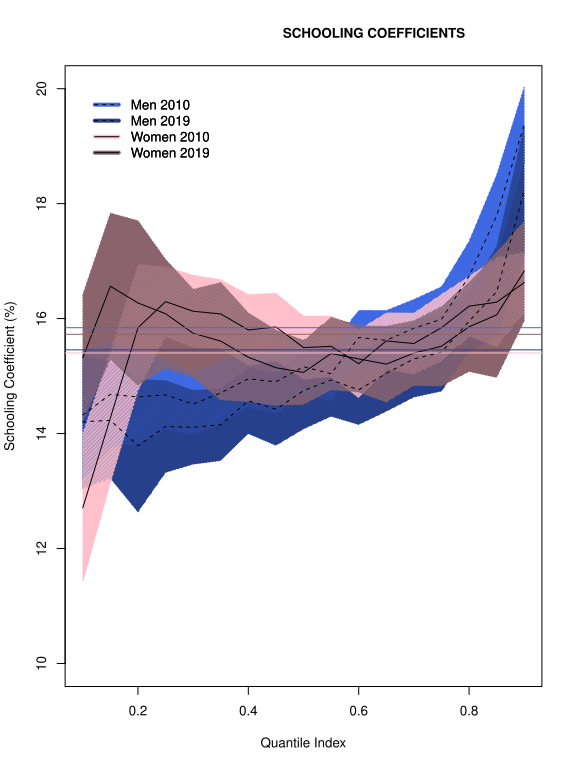
***Figure B1 Figure B2***

1. **Men/Women Comparison**

As an addition to ACFV, this study includes women as well as men for 2010 and 2019 seen in figures C1 and C2. The results are interesting in the lower quantiles, as it seems that women in these quantiles see a larger variation in their income which is driven by additional years of schooling, this almost the same level of variation seen in the highest quantile. As there is little overlapping of the bands at the lower and upper end of the quantiles, these variations are likely to be statistically significant from the returns to schooling for men. On the other hand there is a slight dip in the median neighbourhood as seen in Figure C1, although this looks to be not significantly different to that of men in the same quantile due to the lack of overlap, implying similar estimates.

Compared to their male counterparts, women around the first quartile, experience a larger variation in income explained by their level of education. Despite this Figure C2 shows that earnings for women are still lower in general compared to men.

On the other hand, towards the top of the CQF, men see the largest variation in income as a result of additional years of schooling, and generally have higher earnings than women in these quantiles.



***Figure C1 Figure C2***

1. **Conclusion**

As explored in our results from the expansion of the ACFV paper for 2010 and 2019, we can state that the men’s returns from schooling have significantly improved compared to the early 2000s. We may be experiencing, however, a shift in career options available that might facilitate higher earnings without the need for many years of education. It would be interesting to further expand this analysis in the future and see whether this trend has continued.

Moreover, according to the comparison analysis between the men and women groups, it seems like, in recent years, we are experiencing a reduction in the gender gap for returns of schooling coefficient, even though the higher part of the male distribution seems to be still advantage compared and we can observe a higher wage gap between the two gender in the lower part of the distribution. These results however, may be affected by some selection bias when comparing the two groups as, even though we are controlling for an extensive set of variables, there still may be some underlying characteristics that we did not observe and that would directly affect the estimated results.

As mentioned previously, there is scope for expansion to control for hispanic men and women in the regression. This was not carried out here due to the nature of the data which allowed both observed races to belong to the hispanic group, therefore races have been treated as a subset of whether or not an individual is hispanic. This would require an alternative approach otherwise there is a risk of multicollinearity.