

QUANTILE REGRESSION AS A METHODOLOGY FOR UNDERPINNING PROPORTIONATE UNIVERSALISM

Critical appraisal for PhD by publication

-

ROBERT RUSH

-

March 2019

Abstract

In the social sciences, and beyond, we are often interested in the impact of factors on some outcome. These research questions of interest are traditionally addressed with linear regression, which informs on those factors impacting on the average. Frequently though the interest is not in the 'average' but with those in the tails of the outcome distribution, where for example the low performing or high scoring are contained. This is particularly the case when these analyses are to inform policies to improve on those low performing and the identification and targeting of possible interventions for this. Focusing solely on the average and applying interventions across the board can only widen the gap between those low scoring and better performing.

These traditional modelling methods will not provide information on differential impact of a factor across the distribution and indeed can fail to identify important factors. In addition to the analysis suitable to the research question there are inherent linear regression model assumptions which must be met. To try and address this using traditional techniques by segmenting the data to assess factor impact is inefficient and can have power implications. Also a logistic regression approach provides a cut-point with those on either side, regardless of their proximity to that cut-point being in one group or the other.

Therefore to understand the effect of factors across the outcome distribution we must use different techniques and a quantile regression approach offers an assessment across the outcome distribution and can identify those factors which are influential at different locations on that distribution and is also robust to the assumptions which dog those other traditional methods. Thus with a principled method such as quantile regression analysis, there exists an enormous potential to inform not just basic policy questions, as to relationships amongst factors and outcome, but with the resulting more nuanced answers provide those key policymakers with a more complete evidence base with robust informative estimates on those mediating factors and on who to target.

EXECUTIVE SUMMARY

This report links together a number of publications (Main and Supplementary Outputs). The majority of the publications relate to language development in young children and the way that social risk factors impact upon it. But the report is not about language development as such. Rather it concerns the methodologies by which these relationships are analysed statistically and how the understanding of these techniques more fully informs policy and changes the questions that we are able to ask. Where normally the research question dictates the analysis, awareness of novel techniques allows other, more relevant, nuanced questions to be asked and answered.

Background

In addressing health inequalities/social risk and their impact on later life outcomes, how do we assess potential interventions among these but within a proportionate universalism framework. This framework of support that is proportionate to need, requires us to identify the differentiated needs and also what might mediate the negative relationships. The data used and the analysis here are vital in obtaining robust informative estimates that can inform policymakers of these mediating factors and who to target. Using representative datasets is a step in the right direction but the associated analysis has to be appropriate to this purpose. A quantile regression approach offers an assessment across the outcome distribution and can identify those factors which are influential at different locations on that distribution and is robust to the assumptions which dog those other traditional methods, like traditional regression modelling methods, which, it is argued, are not suited to address those questions of interest. The publications referenced here then are an exemplar of the methods used, both traditional and novel with language and socio-emotional development as the outcomes. Language development in young children has been identified as a key factor underlying school readiness, academic attainment and ultimately employment. Although all children acquire language they do so at very different rates. Clearly there is a strong element of heredity in the way that languages emerge – broadly the children of linguistically able adults tend to have linguistically able children. Yet in any analysis of the language of representative groups of children, it is also clear that social risk factors are strongly associated with the mechanisms associated with the child's language performance. It is these mechanisms and the way that they are conceptualised and analysed which is the focus of this report.

Methods

The report illustrates the issue with traditional modelling methods and contrasts them with a novel method, both with regard to the research questions addressed and also methodologically with specific model assumptions. It extends the work and draws together a series of seven linked peer reviewed papers which discuss aspects of the development of language from birth through to middle childhood and beyond into adulthood. Each paper reports on one of three of the UK birth cohorts, namely the British Cohort Study (BCS70), The Avon Longitudinal Study of Parents and Children (ALSPAC) and the Millennium Cohort Study (MCS). Started in 1970, 1990 and 2000 respectively, these cohorts draw on data from over 40,000 children. The analytic techniques used

include Ordinary Least Squares Regression, Logistic regression, Ordinal Regression and Multinomial regression, and finally quantile regression.

Findings

The more traditional regression techniques, although very flexible in their application, are inherently constrained by the fact that their purpose is to estimate to the average for the population distribution and assume that predictors and outcomes are related in a straightforward manner. But independent factors may be influential at different levels of ability and at different epochs, pointing to the importance of exploring the mediating, moderating and potentially moderating mediating mechanisms across the whole distribution and at different stages. To date such analyses remain relatively uncommon in this subject area. Therefore, it is also important to consider analyses which operate across the distribution of the outcomes. This is especially true in non-normal distributions, a relatively common phenomena in the social sciences and is of particularly relevance to those with a specific interest in groups functioning at the periphery of the distribution such as clinical groups. Interestingly approaches to address these points have rarely been used in the field of child development, much less with regard to oral language skills.

Conclusions

Both Mediation Analyses and Quantile Regression add considerable value to a wide variety of analyses. To date they have been underexploited in the field of child development in general and language in particular. But their utility goes far beyond this particular area and these methods of analysis could readily be introduced as standard practice in a great variety of domains where the predictors are associated with clinical groups of any kind. The clinical groups might involve, but are not confined to, special education, medicine, and the allied health professions, (nursing etc.). In short it is clear from this perspective that basic OLS analyses, which have been the benchmark analytical procedure in the past, do not always present the most comprehensive assessment of relationships or ways to address specific questions, and in fact, could mask important factors.

TABLE OF CONTENTS

			Page
1		Introduction	8
	i.	Data Considerations	9
	ii.	Publications as the exemplar	10
	iii.	Social Risk and Language Development across time	13
2		Regression Models	16
	i.	Ordinary Least Squares (OLS) Regression	16
	ii.	Logistic and Ordinal Regression	16
	iii.	Quantile Regression	17
	iv.	Advantages of Quantile regression	18
3		Mediation Models	23
4		Future Directions	28
5		Conclusions	29
6		Appendices	31
		1. Main Outputs	32
		2. Supplementary Outputs	33
		3. Key Underlying Constructs	35
7		References	40

Conceptualisation of the PhD by Publication

The PhD is conceptualised as drawing together seven peer reviewed publications (See Appendix 2 – Main Outputs) with additional reference to a further ten supplementary Outputs (see Appendix 3 – Supplementary Outputs). The main Outputs follow in temporal sequence except for Main Output #2, for which the work was funded by the English Department for Education in sequence but the specific Output to which reference is made here was accepted for publication some years later).

Responding to University requirements

The aim of the present thesis is not to repeat the content of the included Outputs but to place the work in an overall context of addressing inequalities within interventions and show how different analytical approaches were adopted up to and including the use of a novel methodology. This approach fits with the current thinking around social risk, inequalities and interventions to address these. In so doing, it reflects on the links amongst the papers and integrates the body of work together, commenting on the methodological approaches adopted in the different papers and highlighting strengths and weakness of the different statistical approaches. The Main Outputs are organised historically to demonstrate the progression of the ideas and analytic techniques concerned. The one exception to this is Main Output #2, which was conceptualised in the chronological order and the analysis fed into a report for England's department for Education at the time, but was not accepted for publication as a paper until much later, (2018). In each case, the methods, which are laid out in each of the papers with all the constraints of space that are commonly seen in such formats, are discussed in much greater detail in this document and issues associated with the limitation of a given analytic technique are identified. The document identifies new questions arising from the analyses and culminates by identifying new directions for analyses addressing such questions. The first section of the document begins with an overview of general key issues; sampling, the handling of missing data, the measurement of thresholds – and then details these in the specific main outputs, and also their context; the measurement of language and social and emotional difficulties, the concept of social risk and how these are measured. The second section describes different traditional statistical techniques that have been adopted, and also used in some of the different papers together with their strengths and weaknesses, starting with traditional OLS regression and progressing towards the use of quantile regression.

Finally, it is important to articulate my own contribution to the seven Main Output papers and indeed the Supplementary Outputs included in this report. In each case I was actively engaged in the underlying research as a co-investigator on those funded projects. Of the six Main Outputs, the first (Main Output #1 was funded by the ESRC, Main Output #2, #4 and #5 were funded by the English Department for Education (formerly the Department for Communities, Schools and Families). Main Output#3 was funded by Scotland's Chief Scientist Office (CSO) and Main Outputs #6 and #7 was funded through work carried out for the Australian National Health and Medical Research Council (NHMRC). This has meant being involved in the development of the work, writing appropriate sections of the applications, planning the analysis with key individuals, accessing the relevant cohort datasets, carrying out the analysis and contributing to the interpretation of the results and the write-up procedure. All the statistical analysis was driven by

my own contribution. It is significant that the teams for all of these papers involved a number of different individuals and in such circumstances, it can be difficult to tease out the specific contributions by every individual but in each of the cases presented in the Main Outputs in this report, my leadership of the statistical element of the outputs has always remained distinctive and pervasive.

1. INTRODUCTION

Societal and individual social risk is directly related to health and academic inequalities and these in turn, are linked to lifetime achievement and chances, which are then related to ongoing future social risk through differences type of employment, housing, etc.

There are various ways to address this problem depending on what aspect is being addressed (e.g. universal campaigns aimed at everyone or targeted at particular communities in need). Marmot (2010) labelled the term 'proportionate universalism' to refer to the application of support to all, at a community or individual level, but proportional to the amount of need. This focused on improving the health of socially disadvantaged persons and thereby reducing the difference between those in low and high societal groups. Lowe (2007) proposed a child services specific term of 'progressive universalism' to a universal service 'that is systematically planned and delivered to give a continuum of support according to need at neighbourhood and individual level in order to achieve greater equity of outcomes for all children'. A problem with a universal approach to address differences is that the gap between low and high societal or scoring groups tends to widen with this approach of everyone getting the same thing, due for example, to the uptake of services with those better performing adopting these up in greater numbers. Whereas proportionate universalism is underpinned by a recognition that there is a continuum of need. Therefore, undertaking a proportionate universalism (PU) approach requires an investigation of the associated need continuum and to identify the relative importance of factors on the continuum, which in turn may be amenable to change, in order to positively affect a person's position on it. These factors then being weighted, in accordance, with a PU approach.

Therefore, to undertake this type of support, it is necessary to identify not only factors which influence the outcome (need) but crucially how that relationship is affected across this continuum. Also given a background of social risk, consisting of various factors, it is of particular interest to ascertain which factors mediate, are areas for intervention, in the relationship between social risk and need outcome, and recognising that these may impact differentially across this outcome distribution.

Mediators thus identified as significant in these models may indicate which, where and when possible, interventions may be applied and what magnitude of effect at different locales on the continuum may be achieved. But any policy-introduced interventions must be evidence-based and this is only achieved through thorough evaluation. That 'usual' route to evaluation is to review the existing literature/evidence and then run models on datasets to establish or repudiate those relationships. Several things are key within this process to producing robust findings; namely the sampling and the analyses and their associated facets. Taking each of these in turn:

¹ Throughout this thesis reference is made to **Main Outputs 1-7** which are the seven Outputs which are the main focus of the thesis and **Supplementary Outputs 1-11**. Both groups are listed in full in Appendices 1 and 2 in this document respectively.

i Data Considerations

Sampling

Sampling is central to any scientific enquiry. If we want to draw robust conclusions from the data, the nature of the sampling is of vital importance. But if the sampling is deviant in any way then any results will be compromised. Clinical samples often present challenges when it comes to generalisation because one clinical sample does not often map on to another. The great advantage of representative birth cohorts is that, notionally at least, if they are designed effectively, the sampling will be sound and the conclusions drawn, valid. The study of a cohort which is truly representative of a defined population can offer additional advantages in the estimation of distributions and rates for variables of interest, and their comparison over time or across cohorts (Szklo 1998). Once the assumption of representativeness is met, it is then possible to use inferential statistics with confidence to determine a population's characteristics by directly observing only a sample of the population. But it is important to note that any statistical analysis needs to take into account the different types of sampling. For example, the Millennium Cohort Study is cluster sampled, oversampling particular groups low socio-economic status (SES) and ethnicity, to address the risk of disproportionate drop out. This then requires the use of weighting in the sample, at each wave, and any subsequent analysis needs to account for these weights.

Imputation

Large population based cohorts offer many advantages primarily that their scale (and thus potentially representative nature) makes it possible to generalise more easily and because they allow for more accurate estimates of the relationships of interest. They also have their challenges, primarily that, while it is often possible to recruit a representative sample at birth or when the study is initiated, it is much more difficult to retain membership over time. In practice what often happens is that children (and their families) come in and out of these cohorts at different sweeps. Attrition can be a particular issue if numbers in subgroups fall below what is useful in any analysis and this creates a challenge for the analyst trying to identify a denominator population. Do you take the complete cases only with the result that the sample is often a fraction of the original or do you impute the missing data, modelling gaps using the data that are available? The problem with the former is that any results based solely on the observed/available data may be biased due to characteristics of those remaining in at that point (e.g. low SES groups drop out more than others).

As indicated above, because missing data is such a significant phenomenon in longitudinal population studies, data imputation has been used in the majority of the outputs provided in this document. The results of the observed and imputed analysis that have been reported vary only a little. In some outputs the imputed version of the results is used while in others the observed version was employed. However, in each case a comment is made to the effect that the imputed version is used to check the observed one and the results of the models reported do not change as a function of the imputation. Of course, there would always be a challenge to the

interpretation if there were differences but perhaps because these samples are so relatively large, this is rarely a concern in cohort studies. It is important to stress that caution has been used in administering these techniques to avoid using them if the amount of missing data is disproportionately large. The imputation of clinical outcome data was not undertaken as this would not have been appropriately conservative. Thus, where imputation has been used it has been to backfill missing data to all cases for a more complete sample.

ii. Publications as the exemplar

Referring to the context of the Main Outputs, the content focus of this thesis is of social risk on socio-emotional development and child language. Child language is a domain that has attracted considerable interest over recent years because of the impact that language development has for children, parents and services across childhood and potentially the life course (Supplementary Outputs # 1,2,3,4). There were issues with what research already existed to establish these relationships and the importance of them. No studies existed which had looked at the different types of early language involved in detail (receptive, expressive, pragmatic) with regard to earlier social risk and with social and emotional development. Investigation of those different language types is important to be able to efficiently direct services and any possible future intervention(s). Studies had been undertaken on clinical samples and were not population based. This can introduce a bias which can lead to false conclusions and hence misdirect efforts and funds away from that which matters. Another issue is that the analyses involved in many studies is a linear regression modelling type approach with which to make and assess the strength of relationships amongst social risk, social and emotional development and language performance. The problem here is that these are methods based on predicting the average performance, and the danger is that the distributional differential impact of any resulting covariate could be missed. That is a factor may more strongly affect and be more important to those individuals, say, with low scoring language ability (poorly performing) than it does for those individuals who are higher scoring because of their greater ability, for example.

Social Risk

Social risk experienced in childhood has been shown to negatively influence later adult outcomes (mental health/employment/criminality) (see Main Output #1; Supplementary Outputs #1-4; Snow and Powell 2011). These relationships are dependent on many factors evident within the early years including; educational ability such as language, and social and emotional development. So, the proposed research model was of social risk impacting on childhood language ability and social and emotional development. Therefore, if intervention(s) could be found which improve language ability, then this could affect those outcome prospects in later life and given a relationship between social risk, language and social and emotional development, social and emotional development could also be affected which, in turn, could affect a child's abilities in the early years (Main Output #5).

Child Language

Early studies of clinical groups suggested considerable variability across time – children with poor language abilities in the early years often had poor outcomes but many seemed to grow out of their difficulties (Haynes and Naidoo 1991; Conti-Ramsden et al. 2009) – but these studies commonly drew on clinical samples which were difficult to replicate. Other studies have looked at language development over time in representative populations and, on the whole, found rather different results (Johnson et al. 2010; Whitehouse et al. 2010) but the associated analytical approaches vary and are rarely linked together. Similarly, there are differences in the detail provided on key associated factors such as the child’s social and emotional development (Botting et al. 2016). Nevertheless, it is clear that social risk is implicated in these population models for all aspects of child development (Maggi et al. 2010) if not always in the clinical samples, in part because of the unknown social origins of the samples. In some cases, this relationship, between social risk and child development is reported to be more pronounced than in others (Hart and Risley 1995). It is important then that this wider argument about the nature of language over time is the context in which this dissertation was developed and has been the specific interest of my collaborators on those different analyses. The emphasis throughout is on the development of analytical methods to address issues associated with the relationship between social risk and language development. It is not about language development per se and indeed there are supplementary papers which include relevant analyses to which reference is made when considering the themes below. It is plausible that there are a great many other aspects of child development which could be addressed in this way. Hence, language development is the exemplar, the indicator of the argument rather than the argument itself.

Socio-Emotional Development

It is relatively common in the analyses undertaken, for reference to be made about the child’s emotional and behavioural development. This is not the place to discuss this issue in detail but there is now a general consensus that a child’s communication and it’s behaviour are associated. Of course, this then raises the same questions that were identified for language above concerning the measurement of these skills. A host of such measures, based on both observation and parental/teacher report, have been developed over the years and again, there is the same problem of the phenomenon changing over time. The difference from the point of the birth cohorts is that, while there are a variety of measures available one has come to dominate the field in the UK and Europe and to a lesser extent the US and this is the Strengths and Difficulties Questionnaire (SDQ) (Goodman 1997). The SDQ provides a total score which is the sum of the scores for Emotional, Conduct, Hyperactivity and Peer problems subscales together with a score for the perceived impact of the difficulties experienced. It is important that, although the SDQ is effectively the “industry standard” in that it is used across the literature, it is only a screening test. As such it possesses some of the challenges, of the threshold operationalization of language disorder, to which reference was made above. A result of this is that the associated distribution (as with all screening measures) will have a positive skew, precisely because more differentiation

is required at the bottom of the distribution. Once a child gets over the clinical threshold further questions add no further information.

Sampling

The analyses reported in the Main Outputs utilised data from three cohort studies (BCS70, ALSPAC and MCS). Each of these cohorts has the strength of being population based, commencing at different times (and therefore with different follow up times), having large samples, and containing much relevant data with which to make robust judgments. Appendix 3 contains details of the cohorts and the discussion below raises some issues about the sampling as it relates to each individual cohort.

An issue with the data, given the focus of the work reported here, is the use of the terms “specific language impairment” or “developmental language disorder”, the term used most recently in order to try to correctly characterize children in this respect. Although there are broad identifiers – low language, normal IQ, no major neurological complications etc. – when we come to look at the details of samples in the literature, criteria commonly vary considerably – how low does the child’s language skills need to be, what threshold of “normality” for non-verbal intelligence should be used etc.? It is much easier to model such groups in a population sample and be sure that the children all meet the same criteria. Of course, this raises questions about whether it is possible to generalise across two or more cohort studies which are in themselves representative but are different in time and place. And the answer is that while it may be desirable to do so for a variety of reasons, such an approach needs to be treated with caution.

So, for example the British Cohort Study 1970 (BCS70) “represents” the UK in 1970 but is sampled by taking all those children born in a given week.

iii. Social Risk and Language Development across time

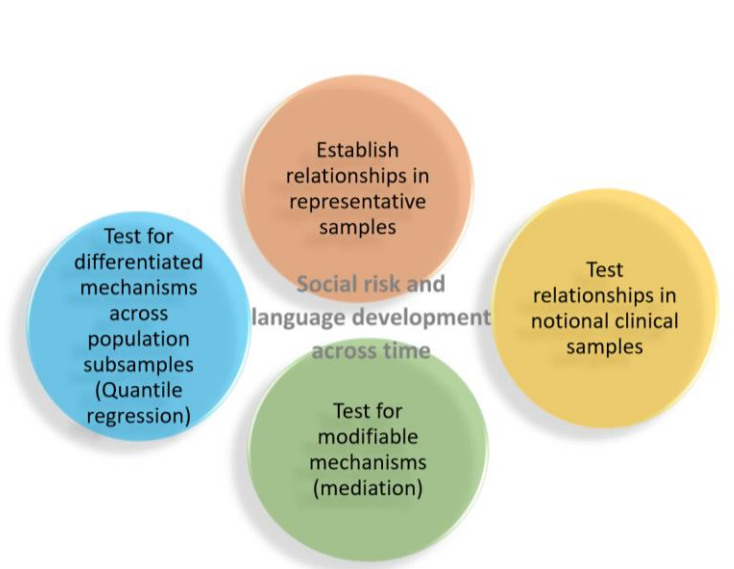


Figure 1. The four main aspects of the thesis progression.

There are four main aspects of the work as represented graphically in Figure 1 above. The first is the need to establish the relationships of interest (social risk and language) in large scale representative samples and across time. The second is to look at these relationships in such populations but modelling the effects in notional clinical samples – i.e. in children with difficulties acquiring language skills – those with Developmental Language Disorders, Language Impairment etc. – to capture how these groups of children function relative to the whole population. The third is to explore the nature of these relationships using more sophisticated statistical models involving so called “third variables” as mediators, where the impact of language skills (and specifically pragmatic skills) as predictors of social and emotional difficulties receives attention. The final aspect is to explore these relationships for different levels of the outcome to see whether the relationships observed in the earlier studies function in a consistent manner or differentially across the distribution. The seven main Outputs link to these four aspects of the work (Main Outputs #1-7).

Establish relationships in representative samples

All of the main Outputs established relationships between their respective indicator(s) of social risk and language across the cohorts and time. These included both expressive and receptive language abilities in children at 2 years (Main Output #2). In addition, Main Output#1 indicated an association of social risk and language with adult outcomes at 34 years, with emotional and behavioural functioning being identified as having been influenced by social risk and early and later language ability also (Main Output #4).

Test Relationships in notional clinical samples

Comparing specific language impairment (SLI), nonspecific language impairment (N-SLI), and typically developing language (TL) groups, the analysis in Main Output #1 indicated that those relationships were consistent for the N-SLI group but rather more mixed for the SLI group. Main Output #3 suggests that, when considering the change in language from 3 to 5 years, maternal education was related to both resilient and increasingly vulnerable groups, but that behavior was only associated with the latter.

Test for modifiable mechanisms (mediation)

Dependent on the aspect of language, Main Output #2 suggests that what parents do with their children is important even against a backdrop of social risk, and this has the potential to inform possible interventions which focus on early language and the home learning environment. Main Output #5 found that there was a partially mediating effect of pragmatic language on the relationship between social risk and adolescent behavior at 13 years, suggesting that interventions targeting pragmatic skills might have the potential to reduce adolescent behavioral symptoms.

Test for differentiated mechanisms across population subsamples (Quantile regression)

Main Output 6 and 7 identified factors that most effectively predict oral language, but specifically revealed the differential effect of those factors across the language distribution, with poverty and early language having a greater effect of those individuals who were low scoring (poorer performing).

Language, however measured, is not a discrete phenomenon. If a child has good language skills, it is often the case that their other skills are also advanced. The relationship between language and other aspects of cognition has been a point of discussion for many years and many authors have argued about the specificity of problems in considering only one of these domains. This applies to language but it is also relevant for speech, for literacy, for numeracy etc.. Some have argued that at the lower end of the distribution, the problem is one of “comorbidity”, but it is probably true to say that the overlap operates to different extents across the distribution. As we shall see below this makes the relative role of both verbal and non-verbal intelligence fundamental in the models discussed in the papers included in this thesis and both are always included in the models, usually with non-verbal skills as a co-variate.

The final point about language which needs to be considered, especially when language across time is being considered is that, unlike the body mass index for example, the nature of language changes over time. Thus receptive and expressive vocabulary in the early years gives way to more complex grammatical structures, narrative and verbal similarities, by the time the child has reached the end of primary school. This means that, on the one hand, at a superordinate level, the construct remains the same but, on the other hand, the specific details of the measures change and it is difficult to know to what extent change scores over time are a function of the child or of the measure.

2. REGRESSION MODELS

i. Ordinary Least Squares regression (OLS) modeling

The most frequently used model to assess the relationship between those factors of interest and a particular outcome is the Ordinary Least Squares (OLS) model. This model estimates the conditional mean and therefore reports on the average relationship between predictors and outcome.

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \dots, n$$

Equation 1. OLS regression equation.

In the above expression, Equation 1, y represents the outcome for n subjects and the modelling produces coefficients, β , for each of the p factors, x , by minimizing the sum of squared errors, Equation 2.

$$\min_{\beta_0, \beta_1, \dots, \beta_p} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2$$

Equation 2. OLS regression minimisation of the sum of squared errors equation.

The OLS regression model has certain important assumptions which have to be met, including; the underlying relationship is linear, homoscedastic and with normally distributed errors. Any violation of these assumptions and the model is void. Transformations of the data can sometimes be used to address the issue of non-linearity but if heteroscedasticity exists and/or the errors are non-normally distributed, this can lead to distrust of the associated standard errors by weighting section(s) of the data. The OLS regression models are reported in terms of model fit statistics and the independent variable coefficients and the associated 95% confidence interval.

This technique is used in most of the primary and secondary Outputs (Main Outputs #2,5,6,7 and Supplementary Outputs #6, 8, 10, 11).

ii. Logistic and Ordinal Regression

Logistic regression operates in a similar fashion to OLS but with a binary outcome. So in this case, the factors of interest predict the probability of that binary outcome, Equation 3, and a natural logarithmic transformation of the coefficients for each factor gives us the associated odds ratio for that factor, Equation 4.

$$\text{logit}(p) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_n x_{in} + \varepsilon_i, \quad i = 1, \dots, n$$

Equation 3. Logistic regression, binary outcome probability equation.

where p is the probability of presence of the outcome event. The model outcome probability is usually classified on a 0.5 split meaning that probabilities above 0.5 from the equation are assigned to the upper binary category and those below are assigned to the lower category.

From the above expression, Equation 3, the logit or logged odds can then be rearranged to give;

$$\rho = \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_n x_{in} + \varepsilon_i}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_n x_{in} + \varepsilon_i}}$$

Equation 4. Logistic regression equation transformed to give probability of outcome.

Unlike OLS modelling estimation which selects parameters that minimises the sum of squared errors, logistic regression selects parameters which maximises the log-likelihood of observing the observed values.

$$L(X|P) = \sum_{i=1, y_i=1}^N \log p(x_i) + \sum_{i=0, y_i=0}^N \log(1 - p(x_i))$$

Equation 5. Logistic regression maximisation of the log-likelihood equation.

Logistic regression requires greater sample sizes than OLS regression and the threshold or cut point always assumes that there are those on either side of the cut who may or may not be accurately characterized by the analysis depending on the measure used and the error involved.

This technique is also used in Main Outputs #1,2, 3 and Supplementary Output # 1,2,3,7,9.

Ordinal regression is an extension to logistic regression and is used when the outcome has ordinal categorical responses of three or more. These response categories can either be predefined, as in from a question in the dataset, or created through cuts in the data. The set of predictor factors are then predicting the odds of a reference category related to the other outcome categories.

Ordinal regression is employed in Main Output #2.

iii. Quantile regression

By definition, a quantile is a position on a distribution corresponding to a particular percentile, a location where a percentage of the data lie below. So the p th percentile is equivalent to the 0. p quantile, and this is the point at which $p\%$ of persons lie below this point. As an example, the

median is the 50th percentile, and the associated quantile =0.5, with 50% of the data having a value below the median. A quantile regression model is similar to other regression models in that it regresses those independent factors onto an outcome. But the difference between OLS, logistic regression, ordinal regression and quantile regression is that the latter allows the investigator to calculate the relationship between independent and dependent variables at pre-specified locations (quantiles) on the outcome distribution. The interpretation of a quantile regression output is similar again to that of the other regressions, with a factor coefficient representing an associated increase in the outcome, but in this instance, it is particular to that quantile, Equation 6. Hence with repeat regressions across the different quantiles of a distribution, we can have a detailed analysis of the differentiated impact of a factor(s).

$$Q_{\tau}(y_i) = \beta_0(\tau) + \beta_1(\tau)x_{i1} + \dots + \beta_{\rho}(\tau)x_{i\rho} + \varepsilon(\tau)_{i\rho}, \quad i = 1, \dots, n$$

Equation 6. Quantile regression equation.

For each quantile level, tau (τ), the solution to the minimisation problem, yields a distinct set of regression coefficients, Equation 7.

$$\min_{\beta_0(\tau), \dots, \beta_{\rho}(\tau)} \sum_{i=1}^n \rho_{\tau} \left(y_i - \beta_0(\tau) - \sum_{j=1}^{\rho} x_{ij} \beta_j(\tau) + \varepsilon(\tau)_{i\rho} \right)$$

Equation 7. Quantile regression minimization of errors equation.

The term “quantile” indicates that these locations are not pre-specified and are selected depending on the hypotheses which the investigator is wishing to explore. Thus, each quantile model produces a distinct set of estimates for the effect at that location for those independent factors.

This method was used in Main Output #6 and 7.

iv. Advantages of Quantile regression

A specific advantage of quantile regression is where our focus of interest is usually in the extreme of a distribution, not in the average. This is often the case in various research areas and for those making policy to address specific groups, where needs require to be identified. Therefore, the problem with using these traditional approaches is that they do not always address the question in hand, especially if the interest is for example, children with low scores on an outcome measure and the conditional mean is not the focus of interest. That is, they are not appropriate to investigate the impact of those factors across the whole distribution. These traditional regression methods operate under an assumption that the relationships amongst the independent factors and the outcome are constant for the different values of the factors. Thus, it may be that a factor is not significant at one, or some, location(s) but is significant at others. Also, the magnitude of a

factor's coefficient estimate, the impact of a factor, could vary for different locations on the distribution and the more quantile models, the more detailed the picture of how factors differentially affect across the range of the outcome distribution. This is most easily seen in pictorial representations (plots) for each factor of their estimates and associated confidence interval across the distribution. Obviously, the more quantiles included in the analyses across the distribution, the more accurate the depiction of the differing effects for those conditioning factors across the distribution.

Therefore, an important consideration when adopting this approach is the relationship between the equivalent OLS regression and the quantile regression. If they do not differ from one another, then one can assume that there is no added value in adopting a quantile approach. Both were run in Main Output #6 for exactly this reason. In each case the OLS regression was also run and allowed a direct comparison of this traditional technique with a quantile regression approach as demonstrable proof of its worth and thus the case for its use, with the results indicating differences across the distribution for some factors. Two examples are presented in the figures below, Figure 2 and 3. In both cases, the quantiles are arranged across along the x-axis and the OLS is represented by the horizontal line in the middle of the diagram. Note that confidence intervals are reported around both the OLS and the quantiles. As can be seen the quantile analysis for naming vocabulary, Figure 2, would suggest that there is a difference for those low scoring, quantile = 0.1, and very little difference after that between the OLS and the quantile, and thus little added value from employing this approach. In this case, it is perhaps not surprising because the predictor and the outcome are different scales of the same standard assessment, albeit tested six years apart, but that effect at low scores would have been missed with the average-based OLS modelling.

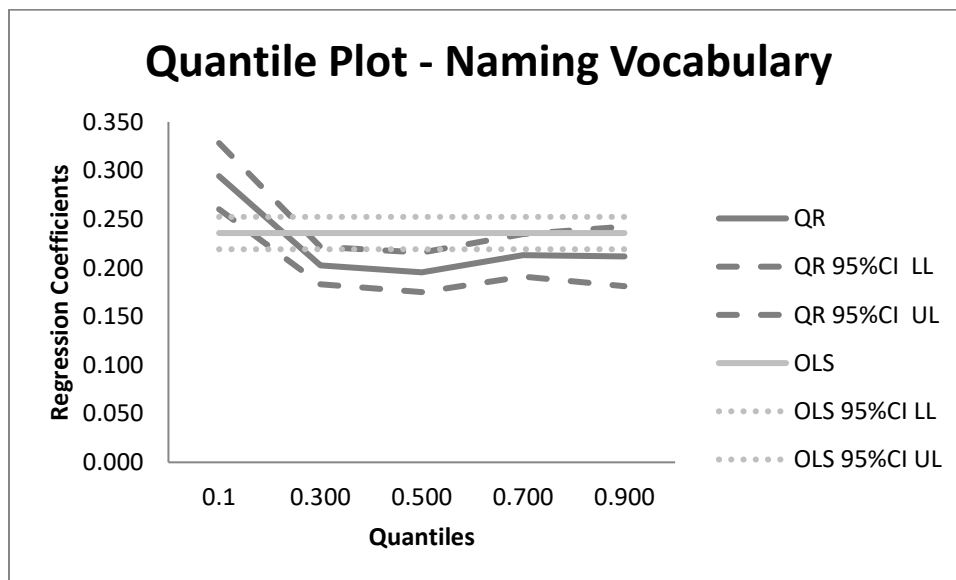


Figure 2. Quantile and OLS regression plot for Naming vocabulary on Oral Language.

By contrast, early book reading paints a rather different picture, Figure 3. Early book reading clearly predicts later language performance but note that it does so much more markedly at the lower end of the distribution of the outcome measure.

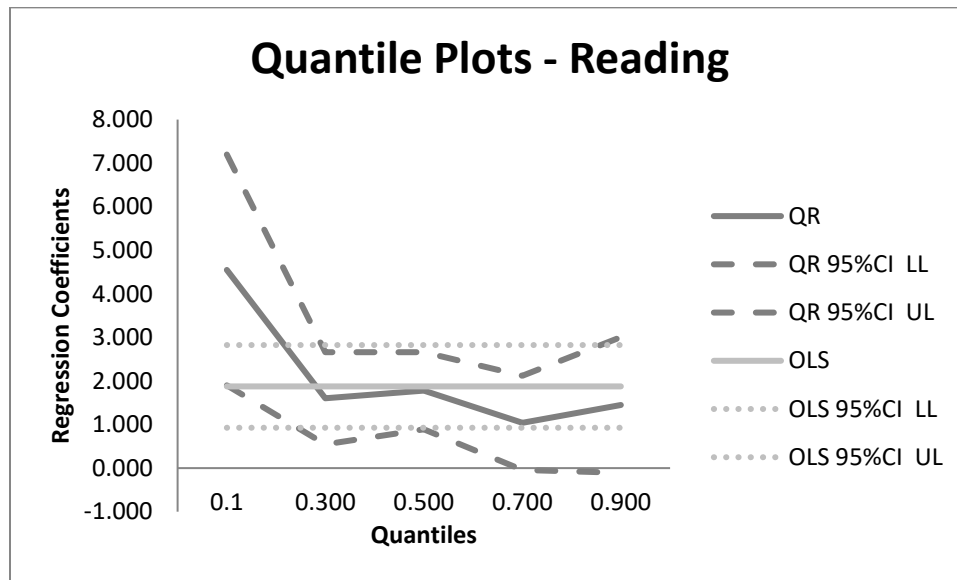


Figure 3. Quantile and OLS regression plot for Reading on Oral Language.

In situations where interest is centred on individuals at the extremes of capability, the usual approach to address this is with logistic regression. There are a few issues, both methodological and analytical, in undertaking this analysis with such cut-offs. While clinical samples obviously focus on the detailed nature of the child's difficulties it is rarely possible to achieve this level of specificity in a population study. There are several reasons for this. The first is simply the time that it takes to carry out full diagnostic work-ups is not practical in large scale studies where detail is often sacrificed for breadth. The second is that large scale populations are commonly and primarily surveys in which parents and carers are asked to speak about the child's performance, needs etc. In such cases, they may or may not be aware of the diagnostic details of any reports on their child and this tends to result in underreporting. When direct assessments are carried out, it tends to be the case that single measures are used at specific time points and so one had to assume that this single measure is sufficient to identify difficulties more broadly. This partly depends on the nature of the test concerned but it also depends on the thresholds on the test that are used. The same issue applies in ordinal regression where the dependent variable is based on 'arbitrary' cuts and this process is open then to the same problems (see Main Output #2).

Taking up the issue of arbitrary cuts, for example, in Main Output #1 the split is at -2 standard deviations, while for Main Output #3 it resides at -1 standard deviations below the mean. So again, the classification is only at one location and does not cover the whole distribution. When the model output captures the observed and predicted responses to determine how well the

model has performed, there is an issue with regard to the numbers and specifically when one outcome category is much smaller than the other, which is always the case when trying to model “clinical” thresholds. Hence, the model always tends to over predict the larger category. Thus, it is harder to accurately assess what is happening in the model. In one of the primary Outputs (Main Output #3), this binary approach was employed to address change between two time points. This led to the creation of a two by two table based on a -1.5 standard deviation cut point. In order to avoid the problem in this case the low scoring groups at two time points was used as the reference group rather than the much larger typically developing group at two time points. In the studies below, notional clinical cut points are used in two of the papers, (Main Outputs #1 and 2) reflecting different approaches to the issue of the threshold. The other Outputs use language as a continuous variable. Regarding findings based on the average regression models, a novel methodology, quantile regression (QR) was employed (Koenker and Basset 1978, Koenker 2005, Hohl 2009, Reeves and Lowe 2009, Koenker 2011, Petscher and Logan 2014, Geraci 2016). Quantile regression assesses the importance of covariates for different quantiles on the conditional distribution (i.e. low scoring/poor language through to high scoring/better language) (Main Output #6, 7).

There are also particular technical advantages to using a quantile regression methodology against those other methods. A major advantage is that whereas OLS modelling, is affected by outliers, requires homoscedasticity across the distribution of the outcome and normally distributed errors, a quantile regression approach is more robust to non-normal errors and outliers. Also, many of the distributions in the social sciences are skewed and have nonlinear relationships with predictor variables, with associated heteroscedasticity. To address this issue with traditional methods, an attempt is made to monotonically transform the outcome to achieve normality. Where a successful transformation can be found, the regression is then modelled on this transformed outcome. However, the inverse transformation of this model does not predict the mean, whereas for quantile regression, as it is invariant to monotonic transformations, it can therefore be used to transform the results back directly to the outcome. One approach to more fully explore the outcome distribution might be to use traditional regression methods on particular sections of the unconditioned distribution (e.g. low scoring, medium scoring and high scoring), in other words, segmenting the sample. But given a normal distribution this segmentation would operate with the associated differing sample sizes and would perhaps be dependent on how those sample segments were arrived at. Again, the quantile regression approach, as it weights on the distance from the quantile(s) of interest, using the whole sample to produce its estimates, does not lose power in the process.

Another feature of quantile analyses is that statistical testing can be undertaken in QR to test the equivalence of factor model estimates across different quantiles (e.g. comparing for difference in estimates between the 0.05 quantile and the 0.5 quantile or whatever the specific comparative quantiles of interest might be). Similarly, statistical comparisons from the estimates can be made to compare changes or differences in shape, scale location and skew.

Although there are several strengths of a quantile regression analysis approach, there are also drawbacks. Using large cohort datasets with bootstrapping of estimates is computationally

demanding and has issues, Chen and Wei 2005, compared with OLS and LR and even more so when adjusting for sampling weights. The latter only being possible currently in R, where to produce the pooled estimates from multiple imputed datasets, the associated syntax can be complex. Other programs can run quantile regressions but do not take in to account any sampling weighting which might be involved in the data design.

The areas in which these models can be applied is expanding as is their use in different types of analyses. In addition to the longitudinal application, survival or time-to-event analyses, characterized by skewed distributions, might be informed by considering the influence of predictors at different locations over their 'survival' period.

Recent studies have highlighted that many areas are important for the routine application of quantile models.

'Quantile regression (QR) has received increasing attention in recent years and applied to wide areas such as investment, finance, economics, medicine and engineering' (Huang et al. 2017).

It has now become clear that the limitations of OLS and logistic regression models are significant and particularly evident where the research question cannot be addressed by traditional methodologies. Quantile regression was originally developed in the field of econometrics where it has been in common use for many years, and it is from here that the use has started to spread to other fields. In child-related research recent work in reading acquisition is evident. Bosch et al, investigating factors influencing reading comprehension, showed from their quantile analyses that vocabulary was related only for 'poor comprehenders' and working memory only for 'good comprehenders' (Bosch et al. 2018). They show that these findings could not have been revealed by the OLS regression that they also conducted. Similarly, Smith et al were interested in blood pressure trends and, in particular, high blood pressure and associates of this. Therefore, they were particularly interested in the upper tail of the blood pressure distribution. They found an association for this involved living in an urban area (Smith et al. 2015) but only in this quantile, a feature that would have been missed had the whole distribution been included in their analysis. Again, utilising health survey data in their investigation of blood pressure drivers, Amugsi and colleagues revealed differential effects for Body Mass Index (BMI) at the 75th quantile, and an increasing effect of age across the distribution (Amugsi et al. 2018). The authors point out that those carrying out interventions should acknowledge those differential effects to increase the effectiveness of their interventions. Using longitudinal data to study BMI in men over time, Bottai et al. found different effects for this at different BMI percentiles, and for activity level and age. The authors maintain that as result of their use of quantiles, they had revealed the specific causes of the obesity epidemic, (Bottai et al. 2014). Fallah et al. (2016) looking at indicators of birthweight with a particular interest in low birthweight, as those babies are much more likely to have health complications found the results from their OLS and quantile regression models in conflict, with gestational age, and weight and educational level of the mother having a linear significant relationship with birthweight. In their conclusions, they recommended that quantile regression models are especially appropriate where the distribution is asymmetric or when there

are outliers (Fallah et al. 2016). Another comparison of the utility of quantile regression compared to OLS, where the focus of interest was in low household income, demonstrated that male education and being a migrant worker were determinants, whereas the associated OLS results were misleading (Pede et al. 2011).

The key features of these studies as far as the quantile analyses is concerned, can be summarised as differential effects of predictors on the outcome distribution, with a particular focus of interest at either end of the distribution, and also the increased capacity for handling non-normal distributions and outliers. The fields of study are varied but these are areas where traditional OLS regression either misinformed or would not be appropriate for the research question in hand. Indeed, for many of the Main and Supplementary Outputs, a quantile approach would have been informative on those factors which differentially influenced behavior or language, across the distribution, adding both more rich information and a more focused understanding of the effects of the model predictors. The traditional methods may, at best, have masked the relative importance of predictor-outcome relationships. However, at worst, as has been shown, traditional approaches may actually misinform the investigator about those relationships. Finally, the QR approach has particular utility where cut points in a distribution are an issue, i.e. in situations requiring the use of logistic models, where differences around the area of interest in the distribution could have been assessed.

While, as has seen above, quantiles have started to be used widely, the approach remains relatively novel in the field of child development in general and indeed within language development in particular. When it was published, Main Output #6 was only the second paper of its type disseminated within the internationally leading journal *Child Development*. The work was developed in response to a call for novel research methods and statistical analysis from the UK's Economic and Social Research Council's secondary data call.

3. Mediation models

The OLS and logistic regression model a linear relationship between the predictor and the outcome and this may be affected by a third factor, which influences the modelling from within either the causal chain or by modulating the effect of one or more of the predictors in the model. These are sometimes referred to as mediators and moderators or sometimes as confounders and interactions, respectively. These terms are important in the analysis of observational data when moving from straightforward associations to putative causal mechanisms.

“If we fail to identify mediators, we are likely to make faulty assumptions about the design of improved treatments.” (Rutter 2009)

The identification of mediators, and by implication potential interventions, is important from a practical point of view. As Rutter says, a mediation analyses investigates a proposed mechanism for a relationship among an outcome and factors of interest in the presence of a potential mediator. In assessing a mediating role, the proposed model is assessed by running OLS or logistic

regressions of the form; outcome on independent and mediator. Initially the total effect is assessed as between the independent and outcome, and then the mediator is added to this regression model. If there is now an indirect path from the independent variable through the mediator to the outcome, then the remaining direct path from independent to outcome will be reduced from the total effect. In fact, the total effect is the sum of the direct and indirect effects. This 'mediation' can be partial with a direct effect still remaining or complete where the direct effect is reduced to zero. An assessment of the size of effect of the mediation can be made by considering the ratio of the indirect to direct path with an associated confidence interval around this estimate.

A moderator is a qualitative (e.g., sex, race, class) or quantitative (e.g., level of reward) variable that affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable. Therefore, it is a variable which has a differential effect on the outcome for an independent – outcome relationship. For example, gender or age may affect the impact of social risk on cognitive outcomes. They can interact with the former to better predict the latter. No assumption is made about causation, simply that one affects the strength of the other. By contrast a mediator is a factor, which as the name suggests, mediates the independent(s) – outcome relationship. That is, the mediator shares the mechanism of the operation of the independent on outcome.

In general, a given variable may be said to function as a mediator to the extent that it accounts for the relation between the predictor and the outcome. Mediators explain how external physical events take on internal psychological significance. Whereas moderator variables specify when certain effects will hold, mediators speak to how or why such effects occur (Barron and Kenny 1986).

From within the studies, the clearest example of the influence of mediator variables is Main Output # 5, in which the role of pragmatic language in mediating the relationship between social risk and childhood behavior has been tested, Figure 4. The question is whether the posited role of pragmatic language reduces the effect of the association between the independent and the dependent variables.

Pragmatics as a potential mediator of the social risk/behaviour relationship

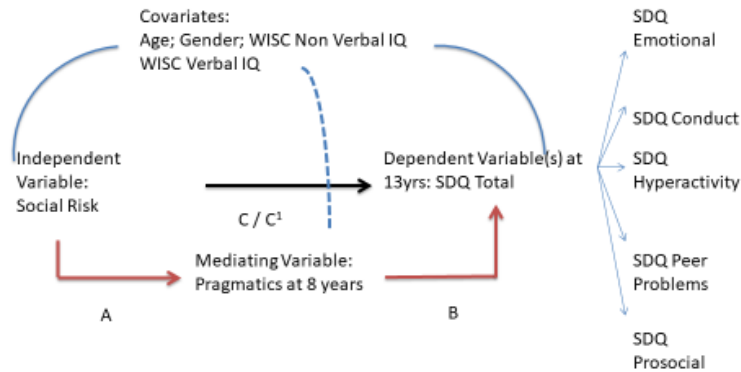


Figure 4. Model schematic for Pragmatics as a mediator for the social risk on behavior model.

Where the proposed hypothesised mediation model under test has to make theoretical ‘sense’ and be backed up by the literature. Also, it is logical that mediation should be a temporal phenomenon, with the mediator coming between the independent and the dependent variable. It makes little sense to test a mediational relationship in a cross-sectional study (with the phenomena of interest measured at the same time point) because it would be impossible to reasonably propose directionality. By contrast, if the outcome is at a later time point and the mediator and independent are earlier then the case for a mediational role is strengthened. As mediation models utilise regression to establish their hypotheses, they are thus prey to the same limitations as those affecting regression methods. Of course, once this type of relationship is established, it is also possible to explore the relationship between multiple mediators involving so called moderated mediation, where the nature of the mediation is influenced by other factors. For example, one might posit that pragmatic skills are the mediator in the above research question but those skills in turn might be sensitive to the effects of gender or age. The moderated mediation is illustrated in Supplementary Output #11.

The means of assessing mediation has changed over the years. Originally 4 steps were outlined, (Baron and Kenny 1986), using traditional regression methodologies, where the relationships among the outcome, mediator(s), and independent(s) are established. In particular, the mediational step, required that a statistically significant relationship existed between the independent and the outcome, such that there was a relationship to mediate in the first place, if there was not, then the mediational analyses went no further than this. This strict adherence to statistical significance was ultimately reviewed for the degree of validity of this approach and currently a study of the magnitude of the coefficient of the independent with outcome is now the driver. This is due to the fact that small effects can be significant in large datasets and

conversely, large effects can appear insignificant in small datasets. So, obviously a coefficient close to zero implies little need to mediate.

Also, to assess the significance and the level of mediation, previously the Sobel test (Sobel 1982) would have been routinely applied. This used an estimate of effect based on the indirect path relative to the direct path, or total effect to direct effect, with the confidence interval around this being based on an assumed normal distribution. Initial steps in mediation (Supplementary Output #8) revealed an issue with values exceeding prescribed critical limits. This issue was taken up in an exchange with David Kenny regarding the interpretation of this and the issue was satisfactorily resolved. This assumption has been relaxed by means of the use of bootstrapping methods. Another change has been in how the mediational analyses are conducted. Originally these were hand calculated by running the associated regressions and by hand calculating the Sobel test and associated confidence intervals. This has now changed through the use of coded syntax to the current macro, outlined by Hayes (Hayes 2009), which also incorporates adjustable bootstrapping for the effect estimates and their confidence intervals to reduce the probability of type 1 errors. The level for the confidence interval can be selected (e.g. 95% or 99%) making it stricter as suits the analysis.

Initially within the Main Outputs (#2), the authors original analyses utilised these methods, involving 4 steps and Sobel test, but on further consideration, the authors ongoing analyses used the bootstrapping methods and did not rely on the 'statistical significance' of the independent with the outcome. Regarding a quantification of the magnitude of a mediating effect, the standardized coefficients for the indirect effect can be used (Preacher and Kelly 2011). The resultant categories for this are based on the square of those of Cohen's proposed "rule of thumb" effect sizes (Kenny 2016); (i.e. .01 small, .09 medium and .25 large).

It could be reasoned that applying a quantile regression approach to a mediational analysis could reveal where, and if, a potential mediating factor might differentially impact on the distribution, and to what level. It is possible that a similar analysis within an OLS or logistic regression framework could identify that a potential mediating effect would not be capable of being revealed, or that its impact would be lessened. Nevertheless, a more thorough assessment across the whole conditional outcome distribution could reveal if, where and to what degree a potential mediator might operate across this distribution. This mediator, if acting as a modifier effect, could then point at a possible intervention target for different loci on the distribution, indicating where the magnitude of intervention might be greater at certain points on the distribution. The author has recently completed this type of analysis using the MCS data (Main Output #7) investigating the mediating role of child naming vocabulary (NV), at 5 years, on the relationships between measures of Child-Parent Relationship (CPR), at 3 years, and Parenting beliefs (PB) on behavior (Strength and Difficulties Questionnaire (SDQ) – at 11 years), Figure 5.

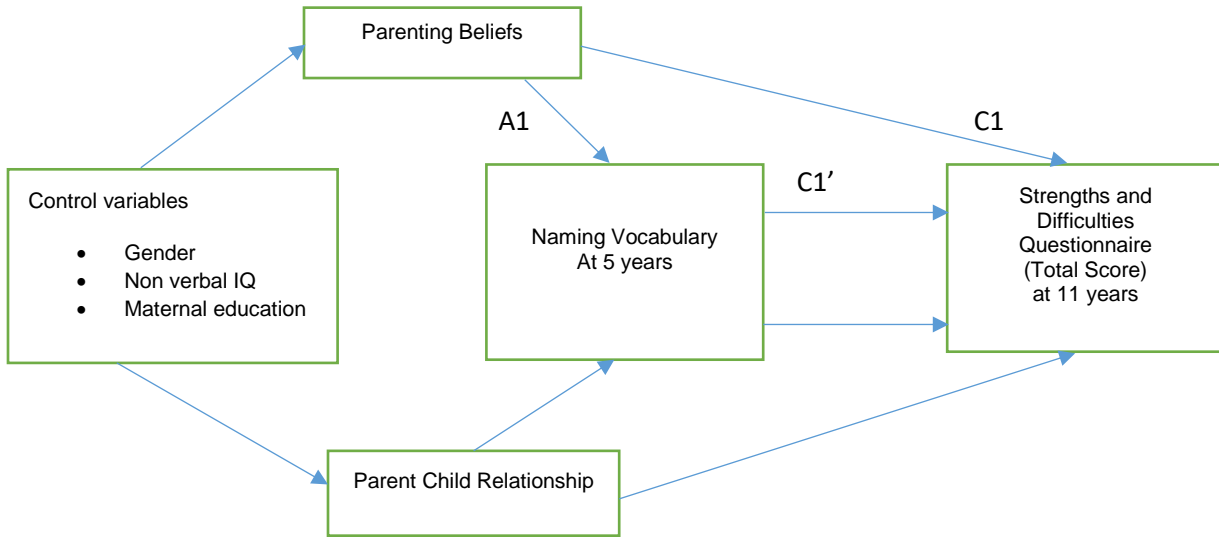


Figure 5. Model for Naming Vocabulary as a mediator for behavior (SDQ) on Parenting Beliefs and Parent Child Relationships.

The analysis within Main Output #7 addresses the research question of whether or not there is a shared mechanism between those independent factors and naming vocabulary on later years behavior. If so how does this then perform across the distribution, while adjusting for IQ, gender and social risk/parental education.

The OLS direct path results established a relationship for CPR but not for PB. Specifically, for the indirect effects, the CPR is significant and of a similar magnitude across the distribution while that for PB is not significant with a reduction in effect from better to poor behaviour (the original magnitude of this coefficient was small as is the resultant change).

$$Q_{\tau}(SDQ_i) = \beta_0(\tau) + \beta_1(\tau)PB_i + \beta_2(\tau)CPR_i, \quad i = 1, \dots, n$$

Equation 8. Quantile regression expression for total effects model, of behavior (SDQ) on Parenting Beliefs and Parent Child Relationships.

$$Q_{\tau}(SDQ_i) = \beta_0'(\tau) + \beta_1'(\tau)PB_i + \beta_2'(\tau)CPR_i + \beta_3(\tau)NV_i, \quad i = 1, \dots, n$$

Equation 9. Quantile regression expression for indirect effects model, of behavior (SDQ) on Parenting Beliefs and Parent Child Relationships, mediated by Naming Vocabulary.

For $\tau = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$. Covariates excluded.

The quantile regression analysis indicated that Child-Parent Relationship had an increasing direct effect from low to high (problematic) behaviour (slight change) and that this remained more or

less of a similar magnitude in the mediated model. The Parental Beliefs estimates were small and constant across the behaviour range with only a slight reduction in the mediated model. It should be mentioned that only the relationships for Parental Beliefs for low scores (quantiles 0.1 and 0.2) were significant, with that for the 0.1 quantile remaining so in the mediated model. Therefore, from the OLS and QR, there is little difference for OLS, and across the quantiles for Parental Beliefs however there is a mediation effect for Naming Vocabulary, with the direct path effect being reduced. The effect for Child Parent Relationship varied across the behaviour range with no alteration in the mediation model. (i.e. the results revealed a differential effect for CPR across the distribution).

More recent research has considered the mediating effect of, child behaviour (SDQ at 5 years), on the relationship between parental mental health and the home learning environment, at birth, on pragmatic, comprehension and expressive language outcomes (at 11 years). The initial complete case analysis, has been completed and the next stage is to impute the missing data and repeat the analyses. These analyses have been undertaken with the Growing Up in Scotland birth cohort.

4. Future Directions

- *Change over time*

Another area of interest which has not been fully thought through yet, is that of using quantile regression to characterise, important factors associated with different amounts of change on a standardized outcome over two time points. In this case, the distribution would be that of change in the outcome, from small through to larger changes. An analysis might then look at the impact of different factors across this distribution and inform on effects over the 'change' distribution. Although quantile regression can handle negative values in the outcome the issue currently unresolved is consideration of all those change possibilities, (increased, remained constant, and decreased). However of consideration is that a person who is performing well and remains so (zero change) is not really of interest and perhaps the solution is, to stratify by initial performance or by removing that group from any analyses, as they would not be a focus of interest dependent on the question being asked. This would be useful as those factors might then act as indicators for possible intervention for those making least improvement but who needed it most, given the details on those who moved from poor to better performing.

5. CONCLUSIONS

If policymakers wish to address the issue of social risk and lifetime achievement then, an earlier consideration of those who are 'worse off' in terms of health or academic performance, is required. Associated studies should reflect this aspiration and not resort to 'average' person analyses employing traditional techniques, which have been shown to be unsuitable for that purpose. A principled analysis must provide robust informative estimates which can then more fully inform those key policymakers, who have been tasked with the development of fully evidence-based policies.

More often than not, the default analysis for many research questions regarding influential factors related to an outcome is the traditional regression approach. While this may be sufficient to satisfy many aims, it lacks full information on those relations. The examples, offered in this overview, of societal and individual social risk and its relation to child health and academic achievement, is seldom interested in the average relationship. These investigations are driven by policies seeking to improve the performance of those lower attaining, and this is the same ambition held within in other fields of study, as previously shown. Applying interventions across the board only widens the gap on performance, and with respect to social risk, it is those who are not at the 'bottom' who improve disproportionately more because of facilitated accessibility in taking up those interventions. Hence a Proportionate Universalism approach of provision of interventional support according to need is a worthy ambition or position on that outcome distribution. In this circumstance, the provision would be calibrated according to a gradient in which those individuals who are the poorest performing would get more, with those better performing, getting least. So the question(s) would appear to be; how to identify those factors which are associated with the various levels of 'performance' across the distribution and the level of association such that, it is known who to target, and to identify the mediating factors and how they perform across the distribution. In so doing, these mediating factors may then point to potential interventions.

A quantile regression analysis approach does address these issues and has the other attractive advantages also of being suitable in cases where regression assumptions are not met. Therefore, it somewhat surprising that it is rarely employed in disciplines outside econometrics. Nevertheless, with the advent of faster computing techniques and awareness this situation is expected to change. Traditional modelling techniques, like ordinary least squares regression, and its derivatives, are all pervasive in the social sciences and can clearly serve a useful function, as far as many questions are concerned. However, they should only ever be seen as a starting point in any analysis. They are never likely, when taken alone, to address those complex pertinent questions in detail which are most often those of greatest interest. Constraining an analysis to the average is incomplete as the relationship of those factors of interest across the outcome distribution remains unknown. Adopting a quantile regression approach provides a thorough assessment of relationships across the whole outcome distribution. In particular, the identification of the differential effect of mediators, by means of the quantile regression approach, allows their impact to be estimated across the distribution, which is particularly informative for future policy impact and allocation of resources. Another benefit of the quantile

methodology is that it is robust to deviations from the underlying assumptions for ordinary least squares (e.g. non-normal distributions), which is a relatively common phenomenon in the social sciences.

As stated, and illustrated, the utility of quantile regression models exceeds that of the exemplars herein and are of use across many domains where the relationship of independent factors to 'clinical group(s)' of any type are the focal point of interest. This consideration is being especially relevant at the lower end of the distribution, where 'clinical' populations rarely function in the way that children do in the middle of the distribution. In scenarios where 'clinical groups' are the focus of research, or as within the overview's exemplars in which research has centred on the capabilities of children scoring at the lower end of the performance distribution, analyses focusing on the middle of the distribution inevitably present inherent challenges. Hence, the rationale for proposing quantile models as a substantive advancement for addressing important clinical questions.

Finally, with a principled method such as quantile regression analysis, there exists an enormous potential to inform not just basic policy questions, as to relationships amongst factors and outcome, but what those relationships look like across the outcome distribution. Thus, awareness of quantile regression can fundamentally change the way in which researchers address these questions and allow and facilitate more relevant, nuanced questions, to be asked and answered.

6. APPENDICES

Appendix 1 – MAIN OUTPUTS

Main Output # 1

LAW, J., RUSH, R., PARSONS, S. and SCHOON, I., 2009. Modelling developmental language difficulties from school entry into adulthood: Literacy, mental health and employment outcomes. *Journal of Speech, Language and Hearing Research*. vol. 52, no. 6, pp. 1401-1416. Available from: [doi: 10.1044/1092-4388\(2009/08-0142](https://doi.org/10.1044/1092-4388(2009/08-0142)

Main Output # 2

LAW, J., RUSH, R., CLEGG, J., PETERS, T., and ROULSTONE, S., (in press). The association of proximal elements of social disadvantage with children’s language development at 2 years: an analysis of data from the Children in Focus (CiF) sample from the ALSPAC birth cohort. *International Journal of Language and Communication Disorders*. vol. 54, no. 3, pp. 362-376. Available from: <https://doi.org/10.1111/1460-6984.12442>

Main Output # 3

LAW, J., RUSH, R., ANANDAN, C., COX, M. and WOOD, R., 2012. Predicting language change between three and five years and its implications for early identification: findings from the Millennium Cohort Study *Pediatrics*. vol. 130, no. 1, pp. 132-137. Available from: <http://www.pediatrics.org/cgi/doi/10.1542/peds.2011-1673>

Main Output # 4

CLEGG, J., LAW, J., RUSH, R., PETERS, T.J., and ROULSTONE, S., 2015. The contribution of early language development to children’s emotional and behavioural functioning at 6 years: an analysis of data from the Children in Focus sample from the ALSPAC birth cohort *Journal of Child Psychology and Psychiatry and Allied Disciplines*. vol. 56, no. 1, pp. 67–75. Available from: [doi: 10.1111/jcpp.12281/](https://doi.org/10.1111/jcpp.12281/)

Main Output # 5

LAW, J., RUSH, R., CLEGG, J., PETERS, T.J. and ROULSTONE, S., 2016. The role of pragmatics in mediating the relationship between social disadvantage and adolescent behaviour *Journal of Developmental and Behavioural Pediatrics*. vol. 6, no. 5, pp. 389–398. Available from: [doi: 10.1097/DBP.000000000000180/](https://doi.org/10.1097/DBP.000000000000180/)

Main Output # 6

LAW, J., RUSH, R., KING, T., WESTRUPP, L., and REILLY, S., 2018. The association between early home activities and oral language skills in middle childhood: A quantile analysis. *Child Development*. vol. 89, no. 1, pp. 295-309. Available from: <http://onlinelibrary.wiley.com/doi/10.1111/cdev.12727/epd>

Main Output # 7

RUSH, R., LAW, J., WESTRUPP, L., and REILLY, S., 2019. Language development at school entry as a mediator of the relationship between early parenting stress and socio-emotional and mental health at 11 years. *Child Development*. (submitted)

Appendix 2 – SUPPLEMENTARY OUTPUTS

Supplementary Output #1

PARSONS, S., SCHOON, I., RUSH, R. and LAW, J., 2009. Long-term Outcomes for Children with Early Language Problems: Beating the Odds. *Children and Society*. Vol. 25, no. 3, pp. 202-214. Available from: [doi:10.1111/j.1099-0860.2009.00274.x](https://doi.org/10.1111/j.1099-0860.2009.00274.x)

Supplementary Output #2

SCHOON, I., PARSONS, S., RUSH, R., and LAW, J., 2010a. Children's language and ability and literacy development: A twenty-nine year follow-up study. *Pediatrics*. vol. 125, no. 3, pp. 459-66. Available from: [doi: 10.1542/peds.2008-2111](https://doi.org/10.1542/peds.2008-2111)

Supplementary Output #3

SCHOON, I., PARSONS, S., RUSH, R., and LAW, J., 2010b. Children's language ability and psychosocial development: A twenty-nine year follow-up study. *Pediatrics*. vol. 126, no. 1, pp. 73-8. Available from: [doi: 10.1542/peds.2009-3282](https://doi.org/10.1542/peds.2009-3282)

Supplementary Output #4

LAW, J. and RUSH, R., 2010. Predicting patterns of development change; language development into primary school in CELSE Abstracts. *Longitudinal and Life Course Studies*. vol. 1, no. 3, page 226.

Supplementary Output #5

LAW, J., MCBEAN, K. and RUSH, R., 2011. Communication skills in a population of primary school aged children raised in an area of pronounced social disadvantage. *International Journal of Language and Communication Disorders*. vol. 46, no. 6, pp. 657-664. Available from: <https://doi.org/10.1111/j.1460-6984.2011.00036.x>

Supplementary Output #6

ROULSTONE, S., LAW, J., RUSH, R., CLEGG, J. and PETERS, T., 2011. *Investigating the role of language in children's early educational outcomes: An analysis of data from the Avon Longitudinal Study of Parents and Children (ALSPAC)*. Monograph - Nottingham: DfE. Available from: <http://www.education.gov.uk/publications/eOrderingDownload/DFE-RR134.pdf>

Supplementary Output #7

LAW, J., RUSH, R., PARSONS, S. and SCHOON, I., 2013. The relationship between gender, receptive language and literacy from school entry through to adulthood: Findings from the BCS70 birth cohort. *International Journal of Speech Language Pathology*. vol. 15, no. 4, pp. 407-15. Available from: doi:[10.3109/17549507.2012.721897](https://doi.org/10.3109/17549507.2012.721897)

Supplementary Output #8

LAW, J., RUSH, R. and MCBEAN, K., 2014. The relative roles played by structural and pragmatic language skills in relation to behaviour in a population of primary school children from socially disadvantaged backgrounds in Special issue of *Emotional and Behavioural Difficulties* entitled "Language, Communication and the child with emotional and behavioural difficulties: implications for practice". Vol. 19, no. 1. Available from: doi:[10.1080/13632752.2013.854960](https://doi.org/10.1080/13632752.2013.854960)

Supplementary Output #9

SPENCER, S., CLEGG, J., STACKHOUSE, J. and RUSH, R., 2016. The contribution of spoken language and socio-economic background to adolescents' educational achievement at age 16 years. *International Journal of Language and Communication Disorders*. vol. 52, no. 2, pp. 184-196. Available from: <https://doi.org/10.1111/1460-6984.12264>

Supplementary Output #10

KING, T., MCKEAN, C., RUSH, R., WESTRUPP, E., REILLY, S. and LAW, J., 2017. Acquisition of maternal education and its relation to single word reading in middle childhood: An analysis of the Millennium Cohort Study. *Merrill Palmer Quarterly*. vol. 63, no. 2, pp. 181-209. Available from: <http://digitalcommons.wayne.edu/mpq/vol63/iss2/2>

Supplementary Output #11

GOODALL, K., RUSH, R., GRUNWALD, L., DARLING, S. and TILIOPOULOS, N., 2015. Attachment as a partial mediator of the relationship between emotional abuse and schizotypy. *Psychiatry Research*. vol. 230, no. 2, pp. 531-536. Available from: <https://doi.org/10.1016/j.psychres.2015.09.050/>

Appendix 3 - KEY UNDERLYING CONSTRUCTS

i) The measurement of language

As indicated above the development of oral language skills is key to a child's progress but it is important to acknowledge a number of features associated with the measurement of language in order to inform the late discussion of the modelling in this thesis. The first is that "language" while it is often used as such is not a unitary construct and can mean a number of different things to the public and researchers. There are a number of ways of separating out its key elements. The first is the difference between oral language other forms of language. As the name suggests oral language represents what the child actually says, but this is different from signed or indeed written language. The second is that a distinction is often drawn between expressive language and receptive language or comprehension, the latter often encapsulating terms such as listening and attention the means by which a child is primed to understand what is said to them. Finally, there are a number of linguistic components which are a number of aspects of language at its most straight forward especially in the early year's vocabulary (the words that a child understands and uses), grammar (the way that those words are combined to form sentences), morphology (the way that words are changed to affect meaning such as plural "s", tense changes etc.) and phonology (the sounds of the language and rules by which they are combined). And finally, we have narrative skills (the ability to combine sentences to make a story, form an argument etc.) and pragmatic skills (the capacity to use language in context – to understand what someone mean by what they say, the ability to understand irony, jokes etc. Many of these constructs can be difficult to assess and often require considerable time to capture effectively. This represents a challenge for large scale birth cohorts where time is often of the essence because of the wide range of behaviours that need to be assessed. As we shall see below different cohorts have addressed these issues in different ways. Often, they rely on single measures of expressive or receptive vocabulary and more recently measures of parental report of vocabulary or pragmatic ability. So these individual measures effectively become a proxy for "language" and therein lies an argument as to how effectively they can serve in this capacity.

In many cases these assessments of language are based on the report of the parent. Measures like the Communicative Development Index require the parent to tick a list of words that the child understands and uses. The Children's Communication Checklist (CCC2) is again a list of questions about the child's communication skills to which the parent or teacher responds. This follows the pattern of most questions in the schedules of the birth cohorts which rely heavily on questionnaires given the parent, in initially in paper form but more recently on computer. The alternative approach is a direct measure of the child's skills by an expert assessor working on behalf of the data collection team of the cohort. Both approaches have their strengths and weaknesses. The former obviously has the advantage that the parent knows the child best and is well placed to give an overall picture including different contexts. But similarly, there is the risk of reporting bias as parents tick a box indicating whether they think understands and/ or says particular word. The latter is more "objective" in the sense that the assessment is independent but especially in younger children runs the risk of underperformance if a child is shy or reluctant for any reason.

ii) The measurement of socio-emotional development

It is relatively common in the analyses undertaken for reference to be made to the child's emotional and behavioural development. This is not the place to discuss this issue in detail but there is now a general consensus that a child's communication and their behavior are associated. Of course, this then raises the same questions as we identified for language above concerning the measurement of these skills. A host of such measures, based on both observation and parental/teacher report, have been developed over the years and again we have the same problem of the phenomenon changing over time. The difference from the point of the birth cohorts is that, while there are a variety of measures available one has come to dominate the field in the UK and Europe and to a lesser extent the US and this is the Strengths and Difficulties Questionnaire (SDQ) (Goodman 1997). The SDQ provides a total score which is the sum of the scores for Emotional, Conduct, Hyperactivity and Peer problems subscales together with a score for the perceived impact of the difficulties experienced. It is important that although the SDQ is effectively the "industry standard" in that it is used across the literature, it is only a screening test and as such has some of the challenges of the threshold operationalization of language disorder to which reference was made above and as a result the distribution of this measure (as with all screening measures) will have a positive skew precisely because more differentiation is required at the bottom of the distribution and once a child gets over the clinical threshold further questions add no further information.

iii) The measurement of "social risk"

Within the literature many terms are used to describe social risk (Index of Multiple Deprivation (IMD), the criteria of the Organisation of Economic Co-operation and Development (OECD), income, housing, social risk (SR), low socio-economic Status (SES), Inequality). They all measure slightly different phenomenon. For example, IMD is measured by postcode area and thus ranks relative deprivation based on different domains within those areas. Each UK country has its own IMD and the, size of area, the number of subdomains and method of combining varies accordingly. (e.g. In England the IMD is a weighted sum of the following domains; income

Employment, education, skills and training deprivation, health deprivation and disability, crime, barriers to housing and services and living environment deprivation). The OECD define an income at sixty per cent below the median as an indicator of those likely to be suffering hardships. Income is commonly measured by equivalised household income. Housing in the UK is commonly measured in terms of type or tenure of housing, overcrowding based on the ratio of rooms per person, and SES is measured using the NS-SEC occupation categorization. In this document we will confine our discussion to social risk in general with specific reference to maternal education.

To operationalize this construct within the different cohorts several of these different indicators at the child's birth were used. Originally with the BCS70 (Main Outputs #1 and Supplementary Outputs #1,2, 3and 7) used a broad approach to social risk employing a number of different indicators; father's and mother's education (education beyond minimum school leaving age), mother had her first child at age 20 or higher, household has no income from paid employment,

social class from the father's occupation (or the mother's occupation, if single): non-manual or skilled manual occupations versus semiskilled or unskilled manual occupations, Housing Conditions at 5 Years of Age, Home ownership, and an overcrowded home. Whereas in the ALSPAC cohort an attempt was made using exploratory factor analytic approach to create a unidimensional construct for social risk, but this resulted in as many factors as items and so ultimately a generic variable of social risk was constructed resulting in a six point scale constructed for the purposes of a similar analysis (from parental occupation, maternal education, housing tenure; overcrowding, financial difficulties, and use of a car) by Schoon and colleagues (Schoon et al. 2004) (see Main Outputs #2, 4,5, and Supplementary Output #6). In the MCS maternal education (GCSE, above and below grade C at 16 years of age) and family poverty (above and below OECD poverty line), Main Outputs #3,6 and Secondary Outputs #4, 10).

These all capture similar overlapping phenomena but their use depends upon their relevance to the argument in question. Thus, the use of different constructs will affect their interpretation. For example, maternal education is often used in the Main Outputs precisely because we understand that maternal education feeds directly into the behavior of the mother and the way that she interacts with her child and thus does or does not foster language and cognitive development in her child (Main Outputs #3,6 and Secondary Outputs #4, 10). This phenomenon has been discussed at length elsewhere in the literature (Dollaghan et al. 1999). Thus, while it is broadly true that they all essentially tapping in to the same underlying construct some are more appropriate in different contexts.

It is common in such analyses to report on social risk as if it is a stable construct. So IMD or OECD is captured at a specific point in the child's development and then it is assumed that it is then constant. This may be de facto a result of the fact that we only have it in a cohort at one time point. But in one the present studies we have also taken this a step further by specifically looking at maternal education as a time varying co-covariate (Supplementary Output #10). This is possible in the MCS where parents were asked at each point whether they had acquired any additional qualifications since they had last been asked to complete a survey questionnaire. The study found that indeed mother reported acquiring additional qualifications at each time point – in fact by the time the child was 11 a third of mothers reported having some additional qualifications. Furthermore this change was associated with higher later language skills, suggesting that change was at least as powerful as early status in such models.

iv) The cohort datasets

The British Cohort Study 1970 (BCS70)

The British Cohort Study takes as its participants 17,196 persons living in Great Britain who were born in 1 week in 1970. Data are available on cohort members at birth, 5, 10, 16, 26, and 30 years,

and most recently in 2004 at 34 years. Demographic information was supplemented by a wide range of parental reports of their child's domestic experience, school reports, tests, and medical examinations.

The Avon Longitudinal Study of Parents and Children (ALSPAC)

ALSPAC is a prospective community ascertained population-based cohort study of all children born to mothers in an area of the west of England in the early 1990s, designed to explore the environmental and genetic factors that affect health and development. All mothers registering their pregnancy within the geographical county of Avon during the period from 1991-1992 were invited to participate. The eligible sample consisted of 20,248 pregnancies and the mothers of 14,541 (71.8% pregnancies were recruited antenatally). Of these 14,541 pregnancies, 14,062 resulted in live births of which 13,988 were alive at one year of age (see Boyd et al. 2013 for a detailed description). The sample was found to have some under-representation of less affluent families and fewer families from black and ethnic minority groups than is the case nationally, although the overall developmental trajectories of the children were similar to national norms for the period (Roulstone et al. 2011).

The Millenium Cohort Study (MCS)

The MCS is a nationally representative cohort of over 18,000 children born in the United Kingdom between September 2000 and January 2002. Informed consent was received

from mothers and partners for participation in the study for themselves and their children and verified at each later wave of data collection. Children's households were sampled randomly from a register of those receiving child benefit, which has an estimated coverage of around 97% of children resident in the United Kingdom. Families were first surveyed at 9 months, when 18,818 children from 18,552 families were contacted (72% of those approached). Families were contacted again when children were aged 3, 5, 7, and 11 years. The MCS oversampled to adjust for specific group drop-out and any analysis has to adjust for this.

v) Imputation

Key to this process is the nature of the missingness – why individuals dropped out in the first place. Indeed there are three types of missingness (Rubin 1976) Missing Completely at Random (MCAR) where the Missing at Random (MAR) and Missing Not at Random (MNAR) . There are many methods for imputing missing data, with some being more principled in statistical terms than others. Methods such as imputing the mean and Last Observation Carried forward are frequently used but these methods underestimate the variance and can bias results. For the outputs here the method of imputation was Multiple Imputation of Chained Equations (MICE) (Van Buuren and Groothuis-Oudshoorn 2011, White et al. 2011) and implemented in STATA

(Royston 2005). This technique effectively imputes those missing data back into the cohort dataset for the purposes of analysis, by modelling those missing data, based on resampling of the observed data and their distribution. The modelling is repeated several times and results in several imputed datasets and then model estimates are averaged across these analyses, with their associated standard errors calculated according to Rubin's rule (Rubin 1987). MICE here operates under the MAR assumption and is able to handle different types of variables (continuous, binary, ordinal and categorical). Although this is a principled method to treat missingness it is vital that the imputation model is correctly specified with all the relevant variables including the outcome, even though this may seem counterintuitive. Not all the variables used in the imputation stage have to be used in the analysis. The resultant imputed dataset should be compared with the observed as should the results of the analyses with only the standard errors being smaller, confidence intervals tighter.

7. REFERENCES

- AMUGSI, D.A., DIMBUENE, Z.T., ASIKI, G., and KYOBUTUNGI, C., 2018. Quantile regression analysis of modifiable and non-modifiable drivers' of blood pressure among urban and rural women in Ghana. *Scientific Reports*. Vol. 8, no. 1, Article number: 8515. Available from: <https://www.nature.com/articles/s41598-018-26991-4>
- AVON LONGITUDINAL STUDY OF PARENTS AND CHILDREN , University of Bristol, Department of Social Medicine, Avon Longitudinal Study of Parents and Children, 2009, *Avon Longitudinal Study of Parents and Children, 1990-2003: Social Science Sampler Datasets*, [data collection], UK Data Service, Accessed 30 June 2019. SN: 6147, <http://doi.org/10.5255/UKDA-SN-6147-1>
- BARON, R.M., and KENNY, D.A., 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*. vol. 51, no. 6, pp. 1173-82. Available from: <http://dx.doi.org/10.1037/0022-3514.51.6.1173>
- BOSCH, L.J. VAN DEN, SEGERS, E., and VERHOEVEN, L., 2018. The Role of Linguistic Diversity in the Prediction of Early Reading Comprehension: A Quantile Regression Approach. *Scientific Studies of Reading*. Vol. 23, no. 3, pp 203-219. Available from: <https://doi.org/10.1080/10888438.2018.1509864>
- BOTTAI, M., FRONGILLO, E.A., SUI, X., O'NEILL, J.R., MCKEOWN, R.E., BURNS, T.L., LIESE, A.D., BLAIR, S.N., PATE, R.R., 2014. Use of quantile regression to investigate the longitudinal association between physical activity and body mass index. *Obesity A Research Journal*. vol. 22, no. 5, pp. 149-56. Available from: <https://doi.org/10.1002/oby.20618>
- BOTTING, N., DURKIN, K., TOSEEB, U., PICKLES, A., and CONTI-RAMSDEN, G., 2016. Emotional health, support, and self-efficacy in young adults with a history of language impairment. *British Journal of Developmental Psychology*. Vol. 34, no. 4, pp. 538-554.
- BOYD, A., GOLDING, J., MACLEOD, J., LAWLOR, D.A., FRASER, A., HENDERSON, J., MOLLOY, L., NESS, A., RING, S. and DAVEY SMITH, G., 2013. Cohort profile: the 'children of the 90s' – the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology*. vol. 42, no. 1, pp. 111-27. Available from: <https://doi.org/10.1093/ije/dys064>
- BRITISH COHORT STUDY, University of London, Institute of Education, Centre for Longitudinal Studies, 2017, *1970 British Cohort Study: Sixteen-Year Head Teacher Questionnaire, 1986*, [data collection], UK Data Service, *4th Edition*, Accessed 30 June 2019. SN: 5225, <http://doi.org/10.5255/UKDA-SN-5225-3>
- CHEN, C., WEI, Y., 2005. Computational Issues for Quantile Regression. *The Indian Journal of Statistics*. vol. 67, Part 2, pp. 399-417.

CONTI-RAMSDEN, G., DURKIN, K., SIMKIN, Z. and KNOX, E., 2009. Specific language impairment and school outcomes. I: Identifying and explaining variability at the end of compulsory education. *International Journal of Language and Communication Disorders*. vol. 44, no. 1, pp. 15-35. Available from: <https://doi.org/10.1080/13682820801921601>

CONNELLY, R. and GAYLE, V., 2017. An investigation of social class inequalities in general cognitive ability in two British birth cohorts. *The British Journal of Sociology*. vol. 70, no. 1, pp. 90-108. Available from: [doi:10.1111/1468-4446.12343](https://doi.org/10.1111/1468-4446.12343)

CRAIG, P., DIEPPE, P., MACINTYRE, S., MICHIE, S., NAZARETH, I., and PETTICREW, M., 2015. Developing and evaluating complex interventions: the new Medical Research Council guidance. *International Journal of Nursing Studies*. Vol. 50, no. 5, pp. 587-592. Available from: <https://doi.org/10.1016/j.ijnurstu.2012.09.010>

DODGE, K.A., PETTIT, G.S., and BATES, J.E., 2008. Socialisation mediators of the relation between socio-economic status and child conduct problems. *Child Development*. vol. 65, no. 2, pp. 649–665.

DOLLAGHAN, C. A., CAMPBELL, T. F., PARADISE, J. L., FELDMAN, H. M., JANOSKY, J. E., PITCAIRN, D. N., and KURS-LASKY, M., 1999. Maternal Education and Measures of Early Speech and Language. *Journal of Speech, Language, and Hearing Research*. Vol. 42, no. 6, pp. 1432-1443. Available from: <https://doi.org/10.1044/jslhr.4206.1432>

FALLAH, R., KAZEMNEJAD, A., ZAYERI, F., and SHOGHLI, A., 2015. Birthweight Related Factors in Northwestern Iran: Using Quantile Regression Method. *Global Journal of Health Science*. vol. 8, no. 7, pp. 116-25. Available from: <https://doi.org/10.5539/gjhs.v8n7p116>

GERACI, M., 2016. Estimation of regression quantiles in complex surveys with data missing at random: An application to birth determinants. *Statistical Methods in Medical Research*. vol. 25, no. 4, pp. 1393–1421. Available from: <https://doi.org/10.1177/0962280213484401>

GOODMAN, R., 1997. The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry*. vol. 38, no. 5, pp. 581-586.

GROWING UP IN SCOTLAND, Scottish Centre for Social Research, 2019, Growing Up in Scotland: Sweep 5 Postcodes, 2009-2010: Secure Access, [data collection], UK Data Service, Accessed 30 June 2019. SN: 7145, <http://doi.org/10.5255/UKDA-SN-7145-1>

HART, B. and RISLEY, T., 1995. *Meaningful Differences in the Everyday Experiences of Young American Children*. Baltimore: Paul Brookes.

HAYES, A.F., 2009. Beyond Baron and Kenny: statistical mediation analysis in the new millennium. *Communication Monographs*. vol. 76, no. 4, pp. 408-420. Available from: [http://dx.doi.org/10.1080/03637750903310360](https://dx.doi.org/10.1080/03637750903310360)

HAYES, A., 2017. *Introduction to Mediation, Moderation, and Conditional Process Analysis. A Regression-Based Approach*. 1st ed., Guilford Press. ISBN 9781462534654

HAYNES, C. and NAIDOO, S., 1991. Children with Specific Speech and Language Impairment. *Clinics in Developmental Medicine*. 119. Cambridge: MacKeith Press.

HOHL, K., 2009. Beyond the average case: The mean focus fallacy of standard linear regression and the use of quantile regression for the social sciences. *SSRN Electronic Journal*. January 2009. Available from: <http://dx.doi.org/10.2139/ssrn.1434418>

JOHNSON, C. J., BEITCHMAN, J. H., and BROWNLIE, E. B., 2010. Twenty-year follow-up of children with and without speech-language impairments: Family, educational, occupational, and quality of life outcomes. *American Journal of Speech-Language Pathology*. vol. 19, no. 1, pp. 51 – 65. Available from: <http://dx.doi.org/10.1044/1058-0360> (2009/08-0083)

HUANG, Q., ZHANG, H., CHEN, J., and HE M., 2017. Quantile Regression Models and Their Applications: A Review. *Journal of Biometrics and Biostatistics*. vol. 8, no. 3, pp. 1-6. Available from: <http://dx.doi.org/10.4172/2155-6180.1000354>

KOENKER, R., 2005. Quantile regression. New York, NY: Cambridge University Press.

KOENKER, R., 2011. Quantreg: Quantile regression. R package version 4.65. Retrieved from <http://cran.r-project.org/web/packages/quantreg/versions/11992>

KOENKER, R., and BASSET, G., 1978. Regression quantiles. *Econometrica*. vol. 46, no. 1, pp. 33 – 50. Available from: <http://dx.doi.org/10.2307/1913643>

LOWE, R., 2007. Facing the future: a review of the role of health visitors. London: Queen's Nursing Institute.

LUMLEY, T., 2011. Survey: Analysis of complex survey samples. R package version 3.24-1. Seattle, WA: Department of Biostatistics, University of Washington

MACKINNON DP, LOCKWOOD CM, HOFFMAN J.M., et al. 2002. A comparison of methods to test the significance of the mediated effect. *Psychological Methods*. vol. 7, no. 1, pp. 83–104.

MAGGI, S., IRWIN, L.J., SIDDIQI, A. and HERTZMAN, C., 2010. The social determinants of early child development: An overview. *Journal of Paediatrics and Child Health*. Vol. 46, no. 11, pp. 627-635.

MARMOT, M., 2010. The Marmot Review. Fair society, healthy lives: Strategic review of health inequalities in England post 2010. London, UK: The Marmot Review.

MARMOT, M., 2015. *The health gap: The challenge of an unequal world*. 1st ed. London, UK: Bloomsbury.

MARMOT, M., FRIEL, S., BELL, R., HOUWELING, T. A., and TAYLOR, S., 2008. Closing the gap in a generation: Health equity through action on the social determinants of health. *Lancet*. vol. 372, no. 9650, pp. 1661–1669.

MILLENNIUM COHORT STUDY. University of London, Institute of Education, Centre for Longitudinal Studies. (2017). *Millennium Cohort Study: Fourth Survey, 2008*. [data collection]. 7th Edition. UK Data Service. SN: 6411. Available from: <http://doi.org/10.5255/UKDA-SN-6411-7>

NATIONAL INSTITUTE OF CHILD HEALTH AND HUMAN DEVELOPMENT EARLY CHILD CARE RESEARCH NETWORK., 2005. Duration and developmental timing of poverty and children's cognitive and social development from birth through third grade. *Child Development*. vol. 76, no. 4, pp. 795–810. Available from: <http://dx.doi.org/10.1111/j.1467-8624.2005.00878.x>

PEDE V, JOYCE LS, THELMA PR, MCKINLEY J., 2011. Determinants of Household Income: A Quantile Regression Approach for Four Rice-Producing Areas in the Philippines. *Asian Journal of Agriculture and Development*. vol. 09, issue 2, pp. 65-76.

PETSCHER, Y., and LOGAN, J. A. R., 2014. Quantile regression in the study of developmental sciences. *Child Development*. vol. 85, no. 3, pp. 861–881. Available from: <http://dx.doi.org/10.1111/cdev.12190>

PREACHER, K.J., HAYES, A.F., 2008. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*. vol. 40, no. 3, pp. 879–891.

REEVES, E. B., and LOWE, J., 2009. Quantile regression: An educational policy research tool. *Southern Rural Sociology*. vol. 24, no. 1, pp. 175–199. Available from: <http://dx.doi.org/10.1111/cdev.12190>

ROULSTONE, S., LAW, J., CLEGG, J., RUSH, R. and PETERS, T.J., 2011. Investigating the role of language in children's early educational outcomes. *Department for Education. Research Report DFE-RR134*.

RUBIN, D.B., 1976. "Inference and missing data". *Biometrika*. vol. 63, no. 3, pp. 581–92. Available from: <http://dx.doi.org/10.1093/biomet/63.3.581>

RUBIN, DONALD B., 1987. *Multiple Imputation for Nonresponse in Surveys*., Chichester, New York. John Wiley and Sons

RUTTER, M., 2009. Understanding and testing risk mechanisms for mental disorders. *Journal of Child Psychology and Psychiatry*. vol. 50, no. 1-2, pp. 44–53.

SCHOON, I., 2006. *Risk and Resilience: Adaptations in Changing Times*. Cambridge University Press.

SCHOON, I., PARSONS, S. and SACKER, A., 2004. Socioeconomic adversity, educational resilience and subsequent levels of adult adaptation. *Journal of Adolescent Research*. vol. 19, no. 4, pp. 383–404.

SMITH, B.L., FUENTES, M., GORDON-LARSEN, P., REICH, J.B., 2015. Quantile regression for mixed models with an application to examine blood pressure trends in China. *The Annals of Applied Statistics*. vol. 9, no. 3, 1226--1246. Available from: <http://dx.doi.org/10.1214/15-AOAS841>

SNOW, P.C., and POWELL, M.B., 2011. Oral language competence in incarcerated young offenders: Links with offending severity, *International Journal of Speech-Language Pathology*. vol. 13, no. 6, pp. 480-489. Available from: <https://doi.org/10.3109/17549507.2011.578661>

SOBEL, M. E., 1982. Asymptotic intervals for indirect effects in structural equations models. In S. Leinhardt (Ed.), *Sociological methodology 1982* (pp.290-312). San Francisco: Jossey-Bass.

SZKLO, M., 1998. Population-based cohort studies. *Epidemiologic Reviews*. vol. 20, no. 1, pp. 81-90.

TRZESNIEWSKI, K. H., and BRENT DONNELLAN, M., 2010. Rethinking “Generation Me”: A Study of Cohort Effects from 1976-2006. *Perspectives on Psychological Science*. Vol. 5, no. 1, pp. 58–75. Available from: <https://doi.org/10.1177%2F1745691609356789>

VAN BUUREN, S., and GROOTHUIS-OUDSHOORN, K., 2011. MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*. vol. 45, no. 3, pp. 1 –67. Available from: <http://www.jstatsoft.org/v45/i03/>

WHITE, I. R., ROYSTON, P., and WOOD, A. M., 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*. vol. 30, no. 3, pp. 377–399. Available from: <https://doi.org/10.1002/sim.4067>

WHITEHOUSE, H.J., WATT, E.A., LINE, E.A. and BISHOP, D.V.M., 2010. Adult psychosocial outcomes of children with specific language impairment, pragmatic language impairment and autism. *International Journal of Language and Communication Disorders*. vol. 44, no. 4, pp. 511–528.