Evidence & Policy • vol 11 • no 1 • 111-24 • © Policy Press 2015 • #EVPOL • #EVPOL
Print ISSN 1744 2648 • Online ISSN 1744 2656 • http://dx.doi.org/10.1332/174426414X14042146202920

practice

A quality assessment tool for non-specialist users of regression analysis

George Argyrous, g.argyrous@unsw.edu.au
Australia and New Zealand School of Government

This paper illustrates the use of a quality assessment tool for regression analysis. It is designed for non-specialist 'consumers' of evidence, such as policy makers. The tool provides a series of questions such consumers of evidence can ask to interrogate regression analysis, and is illustrated with reference to a recent study published in a peer-reviewed journal. The application of the tool highlights the need for non-specialists to develop their critical skills to ensure regression analysis meets methodological norms. They cannot rely on the fact that it has undergone a peer-review process to assume that the evidence is credible.

key words regression analysis • quality assessment • statistical significance • confidence level

Evidence-based decision making often places a great deal of weight on the findings reported in peer-reviewed journals. Lacking expertise in data analysis and reporting, especially when this involves relatively complex techniques such as ordinary least squares (OLS), or linear, regression, 'consumers' of research may rely on the review process in such journals to ensure the results are robust. But is this reliance on the quality control process of peer-reviewed journals too great? Should consumers of research evidence become better skilled at assessing its quality?

Recent assessments of published evidence in peer-reviewed journals suggest that the quality control process is not as thorough as it should be. Steen and Dager (2013), for example, found that a high proportion of published randomised control trials (RCT) and non-RCTs contain serious methodological flaws. Similarly, Song et al (2010) found widespread bias in published studies that favour positive results, and against negative or non-significant results.

Non-specialist users of research evidence, such as policy makers, therefore need their own skills to interrogate the evidence, even when it is published in peer-reviewed journals. It is common for statistics textbooks to detail the 'dos and don'ts' of regression analysis as part of the discussion of its application (for example, Argyrous, 2011). But these methodological norms are not collected together in the form of a tool that a non-specialist can apply. They tend to focus on the needs of those conducting the regression analysis rather than the needs of those using it in decision making. A review of literature, however, finds few tools to help users of research evidence make such an assessment.[1] The only tool for assessing the quality of linear regression analysis this review uncovered was McCloskey and Ziliak (1996, 2004).[2] However, this tool was designed to assess regression analysis in the context of a systematic review of a large number of published papers, and requires more technical knowledge of regression

analysis and inferential statistics than a 'typical' policy maker will possess. For a policy maker, who is usually drawing on a small number of regression analyses in the context of designing policy or programmes, a simpler quality assessment tool is required.

A simpler tool, partly drawing on the more extensive set of questions in McCloskey and Ziliak (1996), is provided in Table 1, which can be used to interrogate linear regression analysis. Such an exercise does not require expertise in quantitative analysis; a basic knowledge of some key concepts and how to 'think critically' is all that is required. We will illustrate how these questions can uncover limitations that can call into question the stated conclusions of regression analysis.

Table 1: Questions to ask of linear regression studies

Rank in importance	Criteria	Assessment scale		
		Yes	No	Insufficient information
1. Essential	• Are the effect sizes large in practical terms (and not just statistically significant)?	☐	☐	☐
	• Where a number of regression models are used, are all the results consistent in terms of the direction and size of the relationship?	☐	☐	☐
2. Important	• Is the explanatory power of the model(s) large enough to justify the use of the model(s) for decision-making?	☐	☐	☐
	• Do confidence intervals for main effect sizes take in values that have different implications for decision making?	☐	☐	☐
	• Were confidence intervals calculated at a confidence level appropriate to the decisions that will be made?	☐	☐	☐
	• Has the model assumed a linear relationship when there is no evidence for such a relationship?	☐	☐	☐
3. Desirable	• Are the results interpreted as association rather than causation?	☐	☐	☐

Table 1 lists these questions in order of importance. This ordering allows readers of regression analysis to deal with situations where some of the criteria are met but not others. This ordering will become clearer when we apply this tool for assessing regression analysis below. Rather than just a simple yes/no choice, we have also built into the quality scale a provision for where insufficient information has been provided by the publication. Without adequate reporting, results should remain provisional, and this scale point will also alert potential users of regression analysis to key questions they can pose to study authors before making a final judgment about the overall quality of the analysis.[3] In the discussion of how to apply the tool we will indicate the reporting standards needed to be able answer each of these questions.

We will use a recent study published in *JAMA Pediatrics*, by Belfort et al (2013), titled 'Infant feeding and childhood cognition at ages 3 and 7 years: effects of breastfeeding

duration and exclusivity', as an illustrative example of how to 'interrogate' research evidence using this tool. This article appeared in the Journal of the American Medical Association, the authors are all eminently qualified in the field, and they work at prestigious institutions such as Harvard Medical School and Boston Children's Hospital. Yet even in this instance, where there is such strong prima-facie evidence that the research is sound, the use of a simple checklist of questions identifies shortcomings that the peer-review process failed to uncover.

The use of regression analysis in the *JAMA Pediatrics* article

The *JAMA Pediatrics* study analyses data from a sample of children, 1224 of who were assessed at age 3, and 1037 at age 7. Breastfeeding duration (the main independent variable for the study) was measured in two ways: the duration in months of any breastfeeding up to 12 months of age, and also the number of months that the child was exclusively breastfed up to 6 months of age.

The main dependent (or outcome) variable was children's cognitive ability, measured by an established set of cognitive tests of intelligence. At 3 years of age, each child was assessed on the Peabody Picture Vocabulary Test (PPVT), and three specific forms of the Wide Range Assessment of Visual Motor Abilities (WRAVMA) test. These latter three were also aggregated into a WRAVMA total score. At 7 years of age, children were assessed on two forms of the WRAVMA test and two forms of the Kaufman Brief Intelligence Test (KBIT). Thus in total there were 9 outcome measures for children's mental abilities.

Linear regression analysis identifies the statistical association between a single dependent variable and a set of independent variables. As we have noted, the main independent variable was duration of breastfeeding, measured in two different ways. The main dependent variable was children's intelligence and cognition, measured in 9 different ways. Thus 18 separate regression models were tested using the data, measuring the correlation between each of the 9 measures of cognition and each of the 2 measures of breastfeeding duration. In each of these 18 regressions, a range of other possible factors – such as maternal age, maternal IQ, annual household income, and parental education level – were also included in the model.

Before proceeding to a detailed application of the tool presented in Table 1 we should note that it assumes a prior question has already been answered in the affirmative. This is the question of *appropriateness*: was ordinary least-squares regression the appropriate technique for analysing these kinds of data to answer this research question? In other words, the checklist in Table 1 presumes that linear regression analysis is suitable for dealing with the problem at hand. We have not included this question of appropriateness in the checklist in Table 1 because it usually requires a level of expertise beyond the scope of a policy maker or other non-specialist consumer of research. But it would not be remiss for such non-specialist consumers of regression analysis to seek out technical experts who might assess the appropriateness of a regression analysis to a particular problem before using the tool in Table 1.

The results of the *JAMA Pediatrics* article

The first task for a critical 'consumer' of evidence is to identify the key elements of the research upon which the main conclusions are based. Research reports are often

brimming with tables, graphs, or numerical calculations, each of which can be assessed for methodological rigour. However, the main conclusions usually rest on a small set of these statistical results, which should be the focus of scrutiny.

Table 4 of the *JAMA Pediatrics* article, reproduced here as Table 2, presents the key results from the 18 regression analyses, which assess the statistical relationship between breastfeeding and cognition. Based on these results, the article concluded with strong policy recommendations 'to promote exclusive breastfeeding through age 6 months and continuation of breastfeeding through at least age 1 year' (Belfort et al, 2013, E8).

Table 2: Adjusted associations of duration of breastfeeding with cognitive test scores

Score	Points (95% CI) per month breastfed	
	Any breastfeeding to age12 months	Exclusive breastfeeding to age 6 months
At age 3 years		
PPVT-III	0.21 (0.03 to 0.38)	0.50 (0.11 to 0.89)
WRAVMA drawing	0.01 (−0.15 to 0.16)	−0.12 (−0.47 to 0.22)
WRAVMA pegboard	0.09 (−0.06 to 0.24)	−0.03 (−0.37 to 0.31)
WRAVMA matching	0.09 (−0.10 to 0.27)	0.00 (−0.42 to 0.41)
WRAVMA total	0.08 (−0.07 to 0.23)	−0.07 (−0.40 to 0.27)
At age 7 years		
KBIT-II verbal	0.35 (0.16 to 0.53)	0.80 (0.38 to 1.22)
KBIT-II nonverbal	0.29 (0.05 to 0.54)	0.58 (0.01 to 1.14)
WRAVMA drawing	−0.08 (−0.33 to 0.18)	−0.05 (−0.62 to 0.53)
WRAML visual memory	0.04 (−0.02 to 0.11)	0.12 (−0.03 to 0.27)

Are the effect sizes large in practical terms?

Regression analysis produces a set of coefficients that quantify any statistical association between a dependent variable and each of the independent variables. In this instance, the regression coefficients measure by how much a one-month difference in breastfeeding is associated with each of the measures of cognition, and these are presented as the numbers in the second and third columns of Table 2 outside the brackets. For example, on the PVT-III test at 3 years of age, each month of breastfeeding is associated with a 0.21 increase in test scores for children who received any breastfeeding up to 12 months, and 0.50 for children who were exclusively breastfed to age 6 months.

These regression coefficients are the *effect sizes* in which we are primarily interested. Any non-zero coefficient signifies that there is some statistical association between the two variables, controlling for all the other demographic variables that are included in the model.

Studies often discuss such effect sizes in terms of statistical significance, as if this is the same thing as practical significance. But it is important to emphasise the limited meaning of the words 'statistical significance': these words suggest that a non-zero

association found in a sample is likely to be due to the relationship being present in the whole population, rather than as a result of random sampling error. And as others have pointed out (for example, Argyrous, 2011, 313–14), even a very small effect can be statistically significant if it comes from a very large sample (as is the case here).

But even where a statistically significant association is reported, we must ask whether it is *practically* significant. In this example, are the changes in intelligence measured by these coefficients 'large enough' for us to describe them as producing important differences in children's cognitive abilities? Failure to extend the discussion of statistical significance to the more important discussion of practical significance has been identified as a common problem with published studies (Kaul and Diamond, 2010).

To facilitate this discussion of practical significance, study authors need to be explicit about the scales used to measure the key variables; the values of the regression coefficients are directly related to how the variables are measured. For example, if one analysed the association between spending on school resources and educational outcome, and expenditure is measured in whole dollars, a coefficient of 0.12 will suggest that for every *one dollar* increase in spending education outcomes will improve by a factor of 0.12. But if expenditure is measured in *thousands of dollars*, using exactly the same data, the coefficient will be 120, indicating that for a $1,000 increase in spending, outcomes will improve by a factor of 120.

This 'order of magnitude' effect may give the impression that a coefficient is relatively large or small, but is simply an artifact of the way that the variables are measured, rather than as a reflection of the actual relationship. Ideally, results will be expressed in terms of units that are relevant to the decision maker. In the example of school expenditure, measuring expenditure in thousands of dollars will usually be more relevant than single dollar values, since decision making is usually conducted in those amounts.

To apply this criterion to the *JAMA Pediatrics* article, Table 3 extracts from Table 2 the largest positive coefficients between breastfeeding and cognitive test scores, which were the focus of the article's discussion.

Table 3: Adjusted associations of duration of breastfeeding with cognitive test scores: Monthly and maximum changes

Score	Any breastfeeding to age12 months		Exclusive breastfeeding to age 6 months	
	Per month	Over 12 months	Per month	Over 6 months
PPVT-III	0.21	2.5	0.50	3.0
KBIT-II verbal	0.35	4.2	0.80	4.8
KBIT-II nonverbal	0.29	3.5	0.58	3.5

The article only discusses the practical significance of its results in a very limited way, and which is possibly an even more egregious error than not discussing it at all, which is to discuss the *maximum possible changes* implied by the subset of positive associations. For example, on the PPVT-III test, for any breastfeeding up to 12 months of age, the difference between children who were *never* breastfed and those breastfed for the *full* 12 months was 2.5 points. The difference in PPVT-III scores between children who

were exclusively breastfed for the full 6 months and those not breastfed at all was 3 points. 'If one is discussing the highest and lowest observed values, it is essential to explain that those values represent upper and lower bounds of a distribution and then include one or more smaller contrasts to illustrate more realistic changes' (Miller and Rodgers, 2008, 126). However, there is no discussion of whether a score difference of 3 points is in any practical sense a meaningful difference in cognitive ability. Moreover, parental decisions around breastfeeding are rarely about a choice between these two extremes; they are usually about durations of a much smaller size, implying much smaller changes in cognitive outcomes.

The best way to assess whether a statistically significant result has any practical significance is to place the results into a decision-making context. For example, we might try to imagine a mother considering whether to continue breastfeeding or not. A mother who is currently performing any breastfeeding, and is deliberating whether to continue for another month or two, may very well consider the 'loss' in cognitive ability of less than 1 point to be not enough to worry about. Similarly, a policy maker, deciding whether health resources should be spent on encouraging mothers to continue breastfeeding or else be spent in some other way, may feel that the size of the effect is not large enough to justify using scarce resources in this way. In other words, even if we accept that there is a statistical association, we may not feel that it is large enough to warrant action in a specific decision-making context.

The meaning of the coefficients is also difficult to interpret when the nature of the scales used to measure cognition are not adequately explained. While the discussion of the coefficients is about the number of test points to which the differences in breastfeeding duration are related, the authors do not inform the reader that these tests have a range from 40 to over 200, with a mean of 100 and standard deviation of 15. Thus even the largest change suggested by all the regression results, which is 4.8 between KBIT-Verbal II and months of exclusive breastfeeding, looks relatively minor when seen in terms of the range and normal variation of this scale. Had the scale ranged from 0–10 with a standard deviation of 3, a coefficient of 0.8 might then be considered very large.

We regard this quality criterion as essential. In other words, regardless of the extent to which a regression analysis meets the other criteria listed in Table 1, failure to find a 'large enough' effect size means the analysis does not provide evidence for any association between the variables upon which it is worth acting.[4]

Were the results consistent?

The previous section focused on the individual coefficients that a regression analysis produces and asks if these are large enough to warrant action. This would be a relatively straightforward assessment if there were only a single regression model being presented. It is not uncommon, however, for multiple models to be developed to measure the same relationship. One reason for multiple models being used to measure the same relationship is because the same broad concepts can be defined in a number of different ways. Using multiple measures for the same broad concepts, such as 'breastfeeding duration' and 'intelligence', is an important means of improving the validity of the results.

The risk of using multiple measures for key variables, however, is that researchers may go on a 'fishing expedition', combining in separate statistical models slightly

different 'versions' of these variables and reporting only those models that produced large effect sizes. Another potential source for constructing multiple models for essentially the same relationship is to use slightly different time periods of data. For example, if one had access to time series data for a particular outcome, measured quarterly for the past 20 years, the precise quarter of data to begin and end the series used for the regression might affect the strength of the results.

To guard against such 'fishing expeditions' it is important that reports present the findings for all the regression models that were tested, not merely those that produced a particular set of results. At the very least, therefore, study authors need to be queried as to whether they have reported all their models or merely a subset.

The *JAMA Pediatrics* research published the results for 18 separate regressions, given that it measured intelligence 9 different ways and breastfeeding duration in two different ways. Table 2 presents the 18 regression coefficients that result (although it is not stated whether these were the entire set of regression results they generated).[5] Do these results all point in the same direction or were there conflicting results among these various measures for intelligence and cognition? A close scrutiny of these coefficients shows that the strength of the relationship varies greatly across the 18 models. For children who received any breastfeeding up to 12 months of age, 3 results seem reasonably strong (PPVT-III, KBIT-II verbal, and KBIT-II nonverbal), but the other 6 are so close to zero as to be negligible; one (WRAVMA drawing) in fact is negative. Noticeably, the effect size for WRAVMA total, which combines the three separate WRAVMA scores, is a very small (0.08).

For children who were exclusively breastfed the results are even more equivocal. Four of the 9 outcomes were *negatively* associated with breastfeeding, and for one other (WRAVMA matching) the association was zero. The study tended to focus its discussion on the 3 outcome measures that were positively associated with breastfeeding (Belfort et al, 2013, E8):

> In summary, our results support a causal relationship of breastfeeding in infancy with receptive language at age 3 and with verbal and nonverbal IQ at school age. These findings support national and international recommendations to promote exclusive breastfeeding through age 6 months and continuation of breastfeeding through at least age 1 year.

But this conclusion can only be reached by ignoring all the other regression results that provide no such support. These other results were not buried away in the paper; they were presented alongside the positive results and required no technical knowledge to assess that they were inconclusive. The only skill required was a preparedness to go back to the source and look for them.

This criterion of quality assessment only applies to studies that model a particular relationship in a number of different ways, rather than a single regression model. Where this criterion is applicable, we regard it as of equal importance to the first criterion that effect sizes be practically significant. Even if the following quality criteria are all met, if some of the regression coefficients are very large in practical terms, while others are effectively zero or negative, it would not be appropriate to use the positive results alone to base decisions.

What was the explanatory power of the whole model?

Regression models are only as good as the amount of variation in the outcome variable that they explain. Regardless of the size of individual coefficients, the explanatory power of the whole model, taking all the independent variables together, needs to be assessed. The extent to which the whole regression model explains the variation in the dependent variable is captured by a statistic called 'R-squared', which ranges from 0 to 1. A value of 0 indicates that the model as a whole cannot explain any variation in an outcome. A value of 1 indicates a model explains 100% of the variation for the dependent variable.

Where reports do provide the value for R-squared, so that we know the percentage of the variance in the outcome explained by the model, they also need to interpret whether it is a 'high' or 'low' value. What percentage of variance explained is 'enough'? There is no universal value between 0 and 1 that applies in all instances; what constitutes 'good explanatory power' is case specific. It is partly a function of the practical uses to which the results will be applied, and partly a function of the explanatory power of similar models in past research. If a policy maker was using the regression results to design programmes that will involve a large amount of public resources, for example, a value for R^2 of 0.8 might be the minimum required. If on the other hand, she was trying to develop a theoretical understanding of a particular set of relationships, and past research has only been able to produce regression models with values for R^2 of 0.2, a new regression model that has a value of 0.4 might be considered very powerful because it explains twice as much of the variation in the dependent variable than past research has been able to achieve.

The regression models in the *JAMA Pediatrics* article bundle up a set of independent variables, such as breastfeeding time, foetal growth and gestational age, demographic variables, and maternal characteristics, and look at the individual effects of these variables on intelligence and cognition. At age 3 and at age 7, children displayed a wide range of scores for each of the cognition tests. How much of this variation is explained by the regression models?

Unfortunately, the article does not provide the R-squared value so that we cannot assess its explanatory power. Without this critical statistic, any conclusions drawn from this study must be heavily qualified. Ideally the authors would have provided the R^2 for each of the 18 regressions, with some accompanying discussion of what would constitute a 'good' value, but this was completely absent from the discussion.

Do the confidence intervals take in a wide range of values?

The *JAMA Pediatric* study, as with many research studies, is based on a sample of children who were followed over time. Sample results may deviate from the values we would get if we could conduct the same study on the whole population. Confidence intervals assess the extent to which this random sampling error may have affected the results.

In other words, when making an inference from a sample to the population from which the sample has been drawn, we need to be aware that the (observed) sample regression coefficients will not necessarily correspond with the (unobserved) population coefficients, and confidence intervals draw our attention to this fact.

Whenever sample results are presented, therefore, we need to ask whether confidence intervals have also been presented,[6] and if so, how wide are these intervals?[7]

Take, for example, the results for PPVT-III in Table 2. The association between this measure of intelligence and each month of breastfeeding is 0.21. However, we cannot dismiss the possibility that for the whole population of breastfed children, the association might be as high as 0.38, *or as low as 0.03*. When we look at all the confidence intervals we see that 12 of the 18 have negative values at their lower end. In other words, when we not only look at the effect sizes, but also allow for the possible range of sampling error, the relationship between breastfeeding and cognition might be positive, negligible, negative, or anything in between! As the paper itself notes when discussing these confidence intervals, albeit in the third last sentence of the paper, 'the lower confidence limits include values with little clinical importance' (Belfort et al, 2013, E8).

While an important quality criterion, the implications of wide confidence intervals are not as critical as the first three we have discussed. Provided that the regression coefficients are sufficiently large, there are no conflicting results, and the regression model has high explanatory power, one can still argue that there is evidence for an association between the variables, even where the confidence intervals are very wide. The effect of confidence intervals that take in a wide range of values is to reduce the strength of the evidence in favour of an association, rather than completely nullifying it.

Have the confidence intervals been calculated at the appropriate level?

The other question a critical reader should ask when presented with information on confidence intervals is whether the confidence level (sometimes also called the significance or alpha level) is high enough, given the importance of the decisions that might be taken on the basis of the evidence. The default level in most research studies, including this one, is 95%. This means that we can be 95% confident that the intervals we construct from the sample results include the real underlying population values. But a critical reader should not always accept this default value.[8] Given the anxiety that decisions about breastfeeding can cause a mother and the rest of a child's family, the important health issues involved, and the public and private resources that are directed toward encouraging breastfeeding, one could argue that we should be 99% confident that the intervals take in the population values for the relationship between breastfeeding and intelligence and cognition. But at a higher confidence level, the intervals become much wider. Thus they are even more likely to stretch into low or negative values, further limiting the main conclusion of the paper.

In other decision-making contexts a policy maker may want a lower confidence level, since there is always a trade-off to be made between two competing types of errors. For example, assume we are testing a programme that requires an increase in school funding but which may lead to improved educational outcomes among school children. Based on the sample of schools in the programme we assess whether educational outcomes have improved sufficiently to justify a rollout of the programme to the whole population of schools. Depending on what we conclude from the sample of schools and what actually is 'true' for the whole population of schools, we could find ourselves making one of two alternative types of errors. These are depicted in Table 4.

Table 4: Possible errors in decision making when drawing an inference from a sample

Decision based on sample	State of the population	
	No relationship between spending and educational outcomes	**There is a relationship between spending and educational outcomes**
Spending does improve educational outcomes	Type 1 error: Consequence is wasted money and resources	Correct decision
Spending does not improve educational outcomes	Correct decision	Type 2 error: Consequence is educational outcomes are not as high as they could be

The types of error that a decision maker could make depend on the actual state of the population for which a decision is being made. It might be the case, that for all schools to which this programme will be targeted, there is no relationship between expenditure and educational outcomes. If in the trial using a sample of schools no relationship was found and the programme not rolled out, then a correct decision is made. However, if a relationship does show up in the sample of schools that are trialling the new programme, possibly because of pure random factors, and therefore the programme is rolled out to all schools, an error is made. The consequence is that money and resources are wasted on a programme that will have no effect on educational outcomes. This is called a Type 1 error: concluding that there is a relationship when in fact there is not.

The other possibility is that there is a relationship between spending on the programme and educational outcomes for the whole population of schools. A strong correlation between spending and outcomes in the trial will lead to the correct decision about the population, and a rollout is justified. However, if in the sample of schools trialling the programme the correlation is not large enough to justify a rollout, then the wrong decision will be made. The programme will not be rolled out to the whole population of schools, denying the educational benefit that would come from it if it was rolled out, and leaving educational outcomes lower than they could be. This is called a Type 2 error: concluding that there is no relationship when in fact there is.

The problem for a decision maker is that reducing the chance of one error increases the chance of making the other error. More specifically, using a lower confidence level (such as 90%) will increase the chance of making a Type 1 error, which in this instance will lead to wasted money and resources. But it will reduce the risk of making a Type 2 error, which in this instance is an untapped opportunity to improve outcomes. Setting a high confidence level (such as 99%) will reduce the chance of making a Type 1 error, but at the expense of increasing the chance of making a Type 2 error.

Which is to be preferred? This can only be determined by the decision maker, assessing in the context in which they are operating, which of the two possible error types should be minimised. Is it worse to waste money or to not introduce a programme that has benefits to students? The resolution of this issue will depend on a range of specific factors such as the tightness of budgets and where children currently sit in terms of educational outcomes. For the issue of breastfeeding, is it worse to promote breastfeeding when it actually does not have any impact on intelligence, or

to fail to promote breastfeeding when it does? At the very least, confidence intervals for a range of confidence levels should be presented so that the reader can decide for themselves what is appropriate in the decision-making context.

Has the model assumed a linear relationship?

Throughout the *JAMA Pediatric* paper, the authors, and subsequent commentators, make statements about the 'per month change' in intelligence and cognition that comes about by breastfeeding, represented by the regression coefficients in Table 2. The discussion assumes that *for each month of breastfeeding* the measures of intelligence and cognition will change at the uniform rate of the effect sizes presented in Table 2. The implication is that measurements of intelligence for the study children were taken at each month. However, the tests of intelligence were administered at only two points: 3 years of age and 7 years of age. The measured changes were then divided by the number of months to generate the 'per month change' values.

It is not always appropriate to extrapolate linear regression results across points in time at which measurements were not actually undertaken. For example, it may be the case that any health benefit from breastfeeding is concentrated in the first few months of a child's life, and this benefit tapers off (or even reverses) after some threshold age. Conversely, the benefit may accrue slowly and then accelerate after a certain age. The practical implications, as regards support structures for new mothers and educational material around the benefits of breastfeeding of either scenario (or others), are very different. But from the information presented, we simply do not know whether the observed relationship between cognition and breastfeeding is linear or not.

Are the results interpreted as association or causation?

As many statistics students are told, 'association is not causation'. Statistical association can be discovered among many data series that have no conceptual underpinning. Regression analysis as such does not prove that such statistical relationships have a causal structure; a more robust study design is required to determine causality. Thus we should always question a study that discusses the relationship between the independent and dependent variables as if causality has been established.

We regard this quality criterion as the least significant. If all the other criteria are met, then a particular regression analysis can still play an important role in evidence-based decision making. Although it does not prove causality, it can nevertheless be strongly suggestive of what action a policy maker might need to take to bring about a desired outcome.

The *JAMA Pediatric* study is very careful not to assert a finding of causality. Instead, the authors claim 'our results *support* a causal relationship of breastfeeding in infancy with receptive language at age 3 and with verbal and nonverbal IQ at school age' (2013, E8, emphasis added). There is a subtle difference between asserting, on the one hand, that causality has been proven, and merely finding *support* for a causal relation, on the other (much like criminal justice cases where the evidence is seen as being *consistent* with criminal activity rather than proving guilt). It is rare for any individual study to conclusively prove causality between two variables such as breastfeeding and cognition, even when a range of possible confounding factors are taken into

account.[9] On this point, the article meets the required standard by not claiming a causal relationship has been proven.

Discussion

A typical policy maker needs to develop skills to assess claims made on the basis of regression analysis. Policy makers should *not* assume that because research is published in high-ranking peer-reviewed journals someone else has checked the quality of the analysis. Evidence-based decision making requires consumers of research to have the critical skills to make such an assessment for themselves. This paper provides a quick tool that requires little technical knowledge of linear regression analysis to assess its quality (assuming that OLS regression is the appropriate form of analysis). By providing a case study of how to make such an evaluation of research quality, this paper hopes to contribute to this process.

There are many other quality criteria we could use to assess a regression analysis. For example, one could investigate whether other assumptions that must hold for linear regression analysis to be valid are met, such as tests for homoscedasticity and the distribution of the error terms. Similarly, one could assess the construct validity of the measures for the particular variables included in the model. For example, the concurrent validity of the WRAVMA test with other widely used tests for cognition has recently been questioned (Obler and Avi-Itzhak, 2011). However, the tool we have developed and applied here focuses on what we regard as the key aspects of regression analysis that a non-specialist is able to assess. For authors of systematic reviews, more detailed assessments of quality will be relevant, but for a policy maker assessing the overall 'relevance' of a particular study, the tool developed in this article should provide sufficient power to discriminate the quality of the evidence.

Notes

[1] A title search of the electronic databases ProQuest, Scopus, and EconLit, using various keyword combinations, was conducted, and the only references found were McCloskey and Ziliak (1996, 2004). A cited reference search was then conducted using these two articles, and this search found no published simple regression analysis tools.

[2] Miller and Rodgers (2008) provide a useful framework for *presenting* quantitative findings, including linear regression. This could be transformed into a tool for assessing such evidence, but it is also focused on the needs of researchers with more technical needs and capacities than the typical policy maker. The following tool for assessing regression nevertheless draws on this article where relevant.

[3] For example, I emailed the authors of the study I use to illustrate the tool below, putting many of these questions to them, but received only one response, asking why I was interested in the information. An email to the journal editor elicited no response.

[4] We note, however, that there may be instances where a finding of no association is actually a cause for action. For example, in evaluating a current programme that presumes a relationship between a policy action and an outcome, a regression study finds no association; this would provide evidence to support an end to the programme.

[5] We assume that these are all the models that were tested, although it would be legitimate to ask whether other measures of the key variables were used but not reported.

[6] Parsons et al (2011) found widespread failure in published studies to report measurement error.

[7] It is worth noting that sometimes 'high' values for R-squared should be questioned, especially where the number of explanatory variables is large relative to the number of cases. The reasons for this and its implications require technical discussion beyond the scope of this paper, but practitioners may refer studies that have very high R-squared values to statistical advisors who can assess whether this issue needs to be considered.

[8] For a more detailed discussion of the factors involved in selecting the confidence level, see Argyrous (2011, 341–3) and Feinstein (1998).

[9] While the study authors themselves were careful not to assert a finding of causality, the coverage this research received in the media was not so restrained. Metro online (2013), for example, reported *Breastfeed your baby till the age of one to boost your child's IQ*.

References

Argyrous, G, 2011, *Statistics for research*, London: Sage

Belfort, MB, Rifas-Shiman, SL, Kleinman, KP, Guthrie, LB, Bellinger, DC, Taveras, EM, Gillman, MW, Oken, E, 2013, Infant feeding and childhood cognition at ages 3 and 7 years: Effects of breastfeeding duration and exclusivity, *JAMA Pediatrics* 167, 9, 836–44

Feinstein, AR, 1998, P-values and confidence intervals: Two sides of the same unsatisfactory coin, *Journal of Clinical Epidemiology* 51, 4, 355–60

Kaul, S, Diamond, GA, 2010, Trial and error: How to avoid commonly encountered limitations of published clinical trials, *Journal of the American College of Cardiologists* 55, 5, 415–27

McCloskey, DN, Ziliak, ST, 1996, The standard error of regression, *Journal of Economic Literature* 34, 97–114

McCloskey, DN, Ziliak, ST, 2004, Size matters: The standard error of regressions, *American Economic Review* 1, 2, 331–58

Metro online, 2013, *Breastfeed your baby till the age of one to boost your child's IQ*, metro.co.uk/2013/07/30/breastfeed-your-baby-till-the-age-of-one-to-boost-your-childs-iq-3903676

Miller, JE, Rodgers, YM, 2008, Economic importance and statistical significance: Guidelines for communicating empirical research, *Feminist Economics* 14, 2, 117–49

Obler, DR, Avi-Itzhak, T, 2011, Concurrent validity of the wide range assessment of visual motor abilities in typically developing children ages 4 to 11 years, *Perceptual and Motor Skills* 113, 2, 377–85

Parsons, NR, Hiskens, N, Price, CL, Achten, J, Costa, ML, 2011, A systematic survey of the quality of research reporting in general orthopaedic journals, *Journal of Bone and Joint Surgery (British Volume)* 93, 9, 1154–9

Song, F, Parekh, S, Hooper, L, Loke, YK, Ryder, J, Sutton, AJ, Hing C, Kwok CS, Pang C, Harvey I, 2010, Dissemination and publication of research findings: An updated review of related biases, *Health Technology Assessment* 4, 8

Steen, RG, Dager, SR, 2013, Evaluating the evidence for evidence-based medicine: Are randomized clinical trials less flawed than other forms of peer-reviewed medical research? *The FASEB Journal* 27, 9, 3430–6