CUBeC

Centre for Understanding Behaviour Change

www.cubec.org.uk

Key Stage 4 Accountability: Progress Measure and Intervention Trigger

Technical Annex: Techniques for producing an unbiased national pupil progress line

Simon Burgess and Dave Thomson

December 2013

Short Policy Report No. 13/11

(Funded by Department for Education)

Centre for Understanding Behaviour Change
Centre for Market and Public Organisation
University of Bristol
2 Priory Road
Bristol
BS8 1TX
UK

www.cubec.org.uk

CUBeC delivers evidence and insight into the drivers of behaviour change to inform and improve policy-making. The Centre combines expertise in a wide range of academic disciplines: economics, psychology, neuroscience, sociology, education, and social research.

Contents

List of figures

List of tables

Summary

The analyses described in this paper have been developed from a range of statistical techniques that could be used as the calculation basis of a pupil progress measure. We show that each of the techniques can be applied to give the basic features required of a pupil progress measure.

The first step in producing a school-level value-added indicator is the construction of a pupil progress measure that meets – or largely meets – certain fundamental statistical principles. A measure of pupil performance that does not meet these characteristics will be unlikely to provide a robust measure of school performance. This paper focuses on the ability of various statistical and non-statistical techniques to provide acceptable measures of pupil progress.

The two smoothing techniques – lowess and kernel regression – are in essence statistical, and with suitable choices of bandwidth we were able to calculate pupil residuals which were substantially unbiased.

Quantile regression – which is used in the US state of Colorado – is a method which makes no assumptions about any underlying distribution of residuals. This basis of this technique provides a ranking of pupil outcomes conditional on their key stage 2 prior attainment. We see attractiveness in this method's ability to be used to develop pupil progress thresholds. However, it is somewhat intensive computationally.

The current DfE method of calculating a measure of pupil progress (from which school value-added scores are calculated) uses Multilevel Modelling (MLM), a refinement of ordinary least squares (OLS) regression. The existing variance components MLM model does not give unbiased pupil residuals across the key stage 2 prior attainment spectrum because of the within and between school clustering of pupils by key stage 2 attainment. We think it essential that, by one method or another, unbiased residuals (or at least substantively so) are a crucial element in measuring pupil progress.

Ordinary least squares regression, kernel regression and lowess smoothing are all suitable techniques for deriving an unbiased national pupil progress line. We also show that a suitable line can be produced by the much simpler method of piecewise regression.

1. Introduction

There were a number of requirements for the value-added measure. The ‘Attainment 8’ key stage 4 points score proposed by the recent secondary school accountability consultation is the outcome. Secondly, only prior (key stage 2) attainment in English and maths could be included as independent variables. Other factors, which research has shown have a bearing on key stage 4 outcomes over and above prior attainment, will not feature in the revised value-addedmeasure.

Typically, value-added measures have compared the attainment of each pupil to the average of all other pupils nationally with the same prior attainment. These differences (residuals) are then averaged at school level. The mean school value-added score is not necessarily a fair and true reflection of the relative benefits pupils receive from their school. Some pupil characteristics – such as ethnicity, special educational needs and free school meal receipt – affect their progress, though not to the same extent across schools. Schools may have made particular provision such that their pupils with these traits made more progress than their peers elsewhere.

In sum, residuals arise from both the general and systematic impact of pupil characteristics other than their key stage 2 prior attainment on their progress, the effects brought about by each school’s pupil profile, and the specific educational policies and practices of each school. This means that school value-added scores (and the scores for pupil groupings) have to be carefully interpreted so that significant and reasonable conclusions can be drawn about school effectiveness.

The initial step in constructing a measure which is fair to all pupils and schools is to ensure that the model of pupil progress given the bases of input and outcome measures delivers residuals which meet desirable characteristics. They should exhibit:

- Monotonically increasing predictions with respect to prior attainment
- Zero mean (that is, unbiased across the key stage 2 prior attainment range)
- Constant variance (‘homoscedasticity’)
- A normal distribution (in order to calculate fair tests of statistical significance at school level)

As we note below, the revised key stage 4 ‘Attainment 8’ measure is still in development. Final decisions on qualification coverage and business rules have not been taken. Nonetheless, GCSE subjects will have greater prominence than before and the range of vocational qualifications included has been severely reduced. For some schools (often but by no means exclusively those with larger proportions of pupils with high key stage 2 prior attainment), the changes will make little difference to curriculum practice: in others the school curriculum will need to be significantly amended if pupils (and the school) are not to be disadvantaged by the proposed value-added measure.

We compare and contrast piecewise, ordinary least squares, multilevel, lowess, kernel and quantile regression. We demonstrate that each method can provide residuals that

are, or substantially are, unbiased. We consider the extent to which residuals are heteroscedastic and normally distributed.

2. Data

We have used a version of the 2012 key stage 4 final dataset provided by DfE. It contains a developmental version of the 'Attainment 8' points score measure proposed by the recent departmental consultation on secondary school accountability.

For the most part, the analysis is conducted on the subset of pupils included in the calculations for the existing 2012 Value Added (VA) measure (INVACALC=1). A number of analyses are performed based on the subset of pupils attending mainstream schools only (INVAMOD=1).

Descriptive statistics

Table 1: Key Stage 4 Outcome Measures 2012

		N	Min.	Max.	Mean	Std. Deviation	Skewness	Kurtosis
Main-stream	Proposed	525,532	.00	580.0	370.2	121.9	-.553	-.110
	Current	525,532	.00	580.0	429.1	91.9	-1.453	3.715
All schools	Proposed	533,062	.00	580.0	365.6	127.3	-.638	.045
	Current	533,062	.00	580.0	424.6	99.5	-1.589	3.799

The proposed outcome measure has more tractable and convenient statistical properties compared to the current 'Attainment 8' measure which make it a more desirable indicator to model. Table 1 shows that the measure is much less 'peaked' (kurtosis is minuscule) and substantially less skewed, but we note that its variance is substantially larger. It is also, from Table 2 below, more strongly correlated with key stage 2 prior attainment (the average of English and maths fine grades) than the current measure.

Table 2: Bivariate correlations (mainstream schools only)

	Current key stage 4	Proposed key stage 4
Key stage 2 (EM)	0.647	0.749
Current Key stage 4		0.871

In Figures 1 and 2 we summarise respectively the current and proposed points scores by key stage 2 prior attainment. Having calculated the mean key stage 2 fine grade in English and mathematics, we have rounded to one decimal place (hence creating a discretised variable). To avoid sparseness of pupil numbers at the lower end of the prior attainment range:

- pupils with a mean fine grade below 1.5 are assigned to 1.5;
- pupils with a mean fine grade between 1.6 and 2.0 are assigned to 2.0;

- pupils with a mean fine grade between 2.1 and 2.5 are assigned to 2.5;
- pupils with a mean fine grade below 2.6 and 2.8 are assigned to 2.8; and
- pupils with a mean fine grade of 5.8 or above are assigned to 5.8.

Linear interpolation has been used in the charts to estimate intermediate values to produce a joined-up line.

Figure 1: Current key stage 4 'Best 8' points (with English and maths bonuses) by prior attainment 2012 (mainstream schools only)

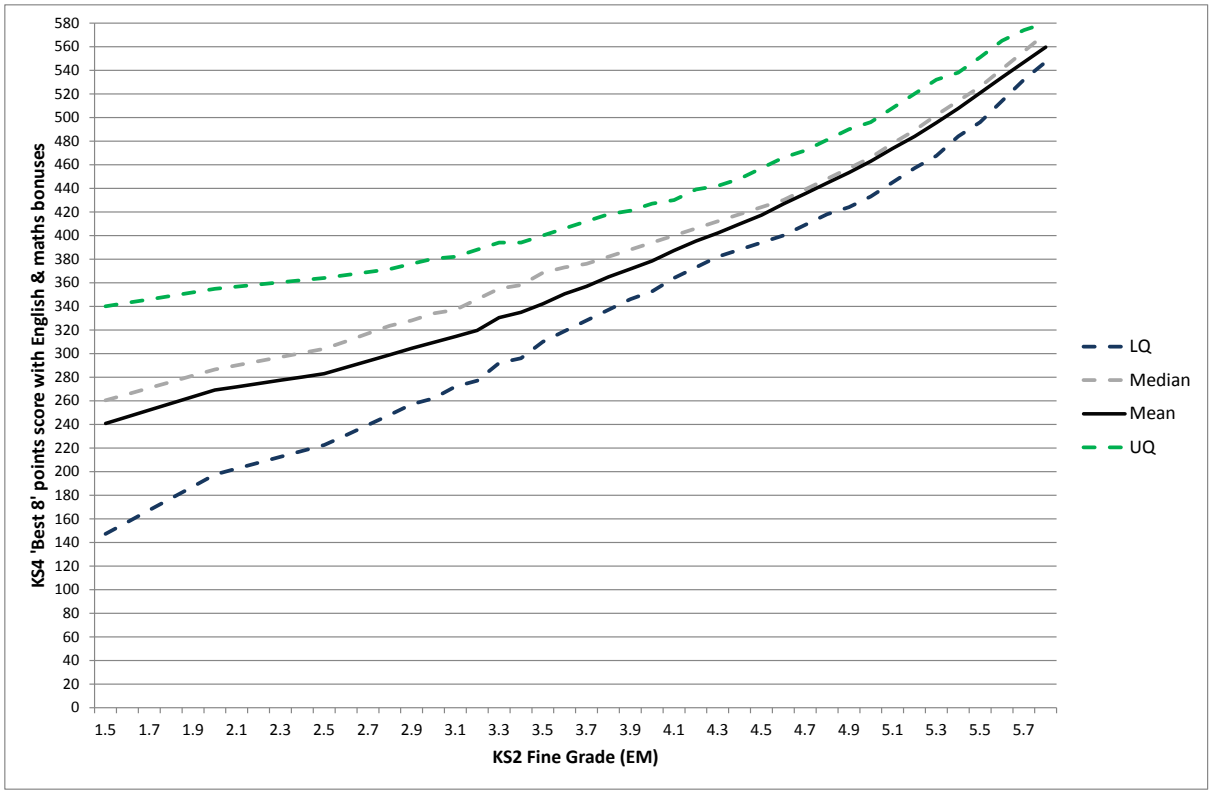
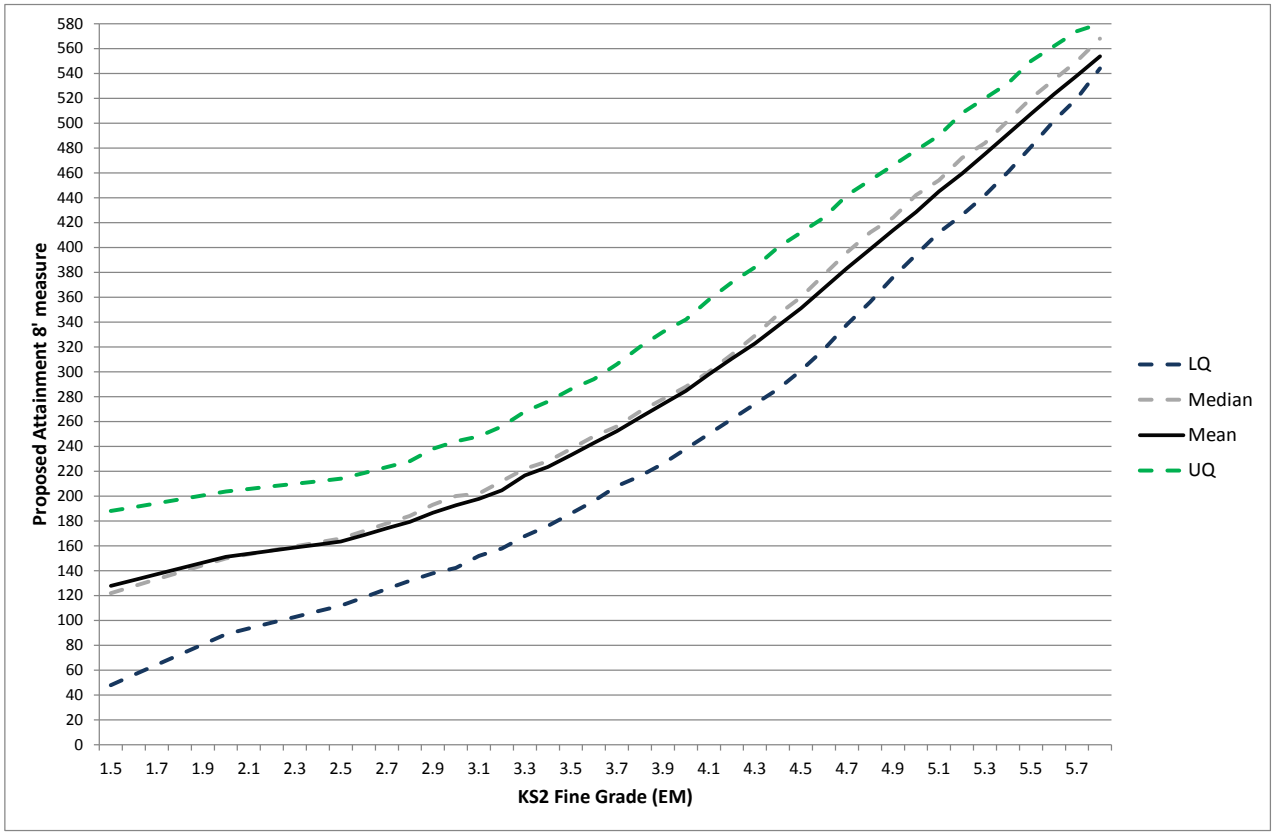


Figure 2: Proposed key stage 4 'Attainment 8' points (with English and maths bonuses) by key stage 2 fine grade 2012 (mainstream schools only)



Compared to the current measure, the proposed measure:

- appears to exhibit more equal variation across most of the key stage 2 prior attainment spectrum;
- increases more sharply as prior attainment rises; and
- has a mean that is generally closer to the median (indicating a more Normal distribution].

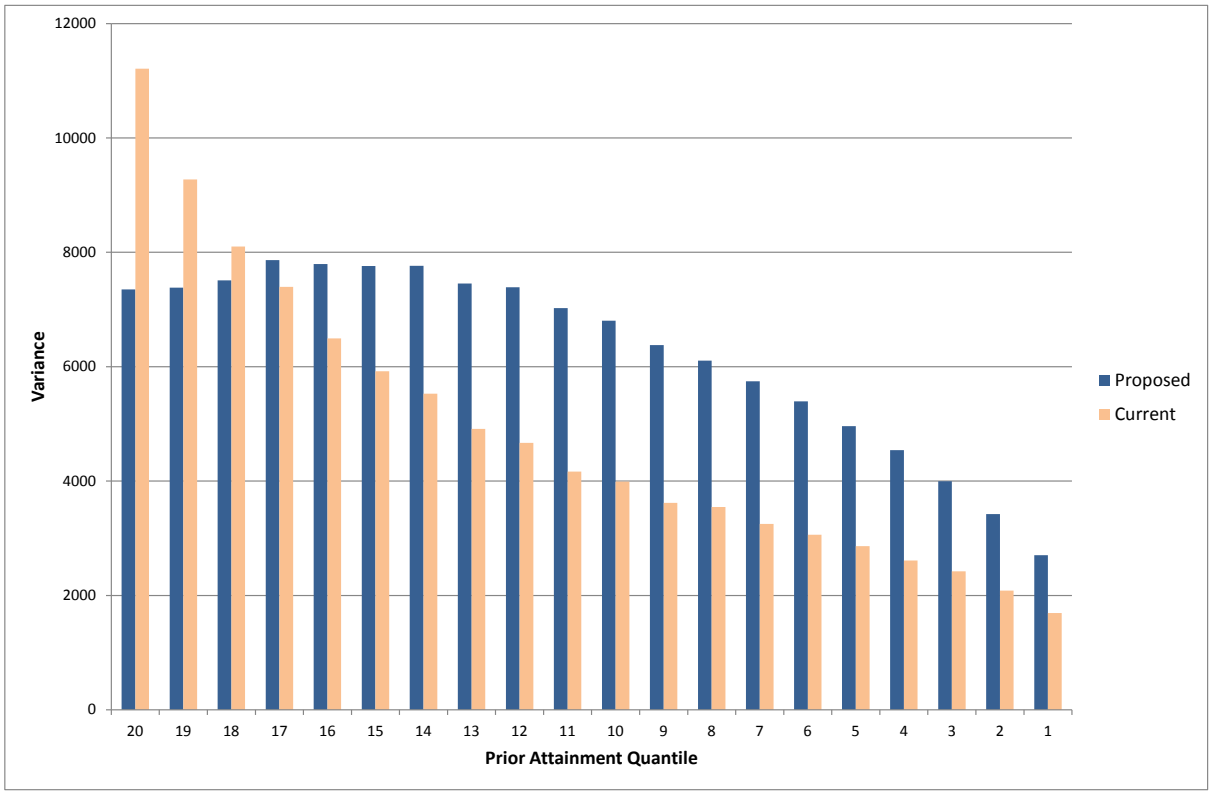
Visually, the mean line of Figure 2 has a more pronounce curvature *within the key stage 2 fine grade range observed* but with a discernible linearity at the upper levels of the key stage 2 and key stage 4 outcomes.

The difference between the lower quartile (LQ) and mean line is around 48 points for the key stage 2 fine grade range from 3.0 to 4.7 inclusive. This range covers over half of pupils nationally in mainstream schools. The gaps are wider at the lower end, and narrower at the top end. These will be important features if setting a 'minimum expected progress level' for pupils relative to the national line.

As is suggested by Figures 1 and 2, the proposed measure is also less heteroscedastic than the current measure. This will be a key consideration in the calculation of school-level scores, particularly for those schools with disproportionately large cohorts of low-ability or high-ability pupils.

Figure 3 below shows the variation in key stage 4 outcomes on the current and revised measures conditional on prior attainment quantile. Table 1 showed that the revised measure had a much wider spread than the current measure and this is reflected in this chart. But we observe that, unlike the current measure, variance is more distinctly constant for pupils with below average levels of prior attainment.

Figure 3: Variance in current and proposed key stage 4 point scores 2012 by prior attainment quantile (mainstream schools only)



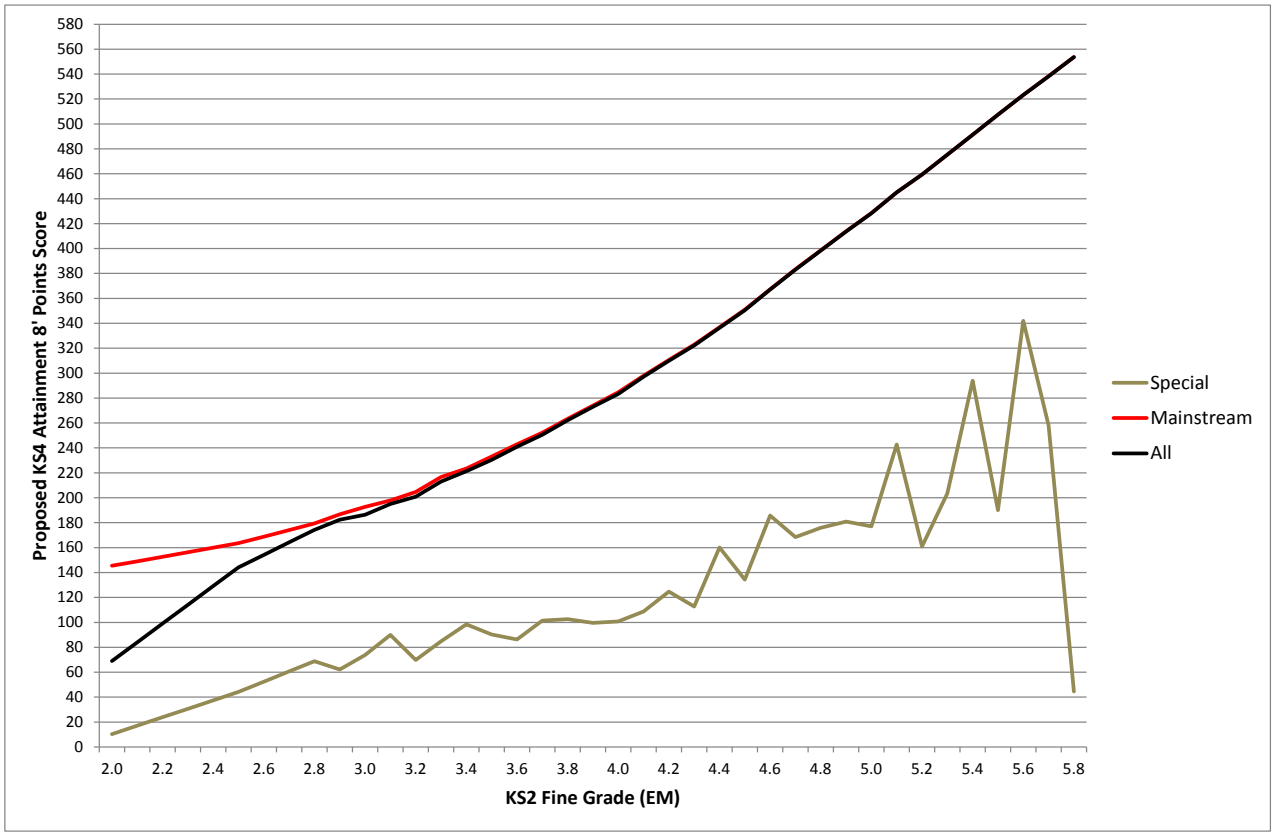
The usual method of calculating the significance of a school value-added score is to use the national variance unexplained by the value-added model. We can see that if the general shape of that distribution were to hold after the value-added model had been tuned to give unbiased residuals, then schools with relatively more higher attaining key stage 2 pupils could be less likely to be shown as having a value-added score that was significantly different from average (and vice versa).

Further, if schools were set an absolute threshold of pupil progress in terms of key stage 4 point scores, those with greater proportions of higher key stage 2 achieving pupils might be considered to have passed a 'progress threshold test'. Of course, key stage 4 scores are upper-bounded (at 580 points) and this has an impact on the residual variance.

Any school value-added score must be accompanied by confidence intervals (CI). Since the CI depends on the size of the unexplained variance, it follows that any substantive departure from homoscedasticity is important. The rescaling of the key stage 4 grade distribution can substantially improve this desirable characteristic.

The current value-added model was based on pupils in mainstream schools only, primarily because the fitting of school lines for pupil in special schools in the MLM structure disrupted the fixed part of the model. We seek to include pupils in special schools in the model as a basic premise. However, as Figure 4 below shows, this will have a significant bearing on the position of the lower end of the national line and, therefore, the value-added scores of schools with disproportionately large numbers of pupils with low prior attainment, most of which will have scores which are significantly below average. We adopt a modelling strategy that creates both 'mainstream only' (_M) and 'all pupils' (_A) versions of models.

Figure 4: Mean proposed key stage 4 'Attainment 8' points (with English and maths bonuses) by key stage 2 fine grade 2012 (all pupils)



3. Fairness to pupils

The current MLM-based value-added model produces predictions that yield small, but non-trivial, biases in value added scores with respect to prior attainment. In order to demonstrate that any alternative model structure produces unbiased residuals, we first divide pupils into 20 quantiles based on finely graded key stage 2 points and examine the mean residual and residual variance in each.

Given the factors which can be included in the value-added model, we judge that a model can be said to be 'fair' to all pupils if:

- The mean residual for each prior attainment quantile is zero
- There is constant residual variance between quantiles
- There is a similar distribution of residuals between quantiles

To minimise the risk of schools not seeking to achieve the best outcomes for all their pupils, we consider that all pupils with both key stage 2 and key stage 4 results should be included in the value-added model.

We note that about 5% of maintained school pupils who reach the end of key stage 4 are not included in value-added models because they do not have prior attainment data. For the most part, these pupils enter schools (usually from overseas) during their secondary education. Such pupils are not randomly distributed amongst schools: for instance, the 2012 value-added measure at eight schools excluded 40 -50% of their pupils because they had not been assessed at key stage 2.

Although we do not advise the imputation of key stage 2 results for these pupils for the value-added model, we do think it worthwhile to consider how 'predictions' of key stage 4 performance (which we cover generally in another paper) might be constructed. These could, for example, be structured around the length of time in education the pupils have spent in English schools and whether English is their mother tongue.

4. The simple model

The simplest model is to use piecewise regression and create a set of dummy variables for key stage 2 fine grade (as shown in Figures 1 and 2) and entering these into a regression model. This is akin to producing a transition table showing the mean key stage 4 points score for each discrete value of key stage 2 fine grade. It would produce the mean line shown in Figure 2 without any interpolated values.

This method yields unbiased residuals. The mean residual is exactly zero for each value of mean key stage 2 fine grade shown in Figure 2.

In order to compare various models more generally, we divide pupils into one of 20 approximately even sized bands based on prior attainment. In Table 3, we show the mean value-added score from the simple model, together with 95% confidence intervals. Mean residuals for the highest group and group 10 are significantly different from zero-although the degree of bias is still relatively small. These biases are caused by rounding finely graded points scores (e.g. 24.6) into fine grades (e.g. 4.1).

Table 3: Mean residuals by prior attainment quantile, simple model (mainstream schools only)

Prior Attainment Quantile	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum
				Lower Bound	Upper Bound		
20	20,767	-.08	84.18	-1.23	1.06	-197.84	388.68
19	26,212	-.25	84.48	-1.27	.78	-233.06	351.40
18	26,280	-.45	86.19	-1.49	.60	-263.37	318.74
17	25,780	.20	87.82	-.87	1.28	-284.68	298.87
16	26,414	-.27	88.81	-1.34	.81	-297.96	282.04
15	27,197	.55	87.94	-.49	1.60	-322.88	269.40
14	26,646	-.85	87.83	-1.91	.20	-336.90	257.12
13	26,872	-.12	86.68	-1.16	.91	-351.05	243.10
12	26,268	.56	86.36	-.49	1.60	-367.33	228.95
11	27,074	.66	83.82	-.34	1.66	-367.33	212.67
10	26,284	-1.34	82.59	-2.34	-.34	-383.34	196.66
9	26,073	-.70	80.16	-1.67	.27	-398.47	196.66
8	26,776	-.17	78.67	-1.12	.77	-413.84	181.53
7	26,209	.04	75.94	-.88	.96	-428.52	166.16
6	27,080	.70	73.53	-.17	1.58	-445.06	151.48
5	26,360	-.04	70.37	-.89	.81	-459.45	134.94
4	25,508	.08	67.96	-.76	.91	-475.34	120.55
3	27,638	-.02	63.24	-.77	.72	-491.31	104.66
2	26,317	.29	58.66	-.42	1.00	-507.44	88.69
1	26,804	1.13	50.96	.52	1.74	-538.28	72.56
Total	**524,559**	**.00**	**78.95**	**-.21**	**.21**	**-538.28**	**388.68**

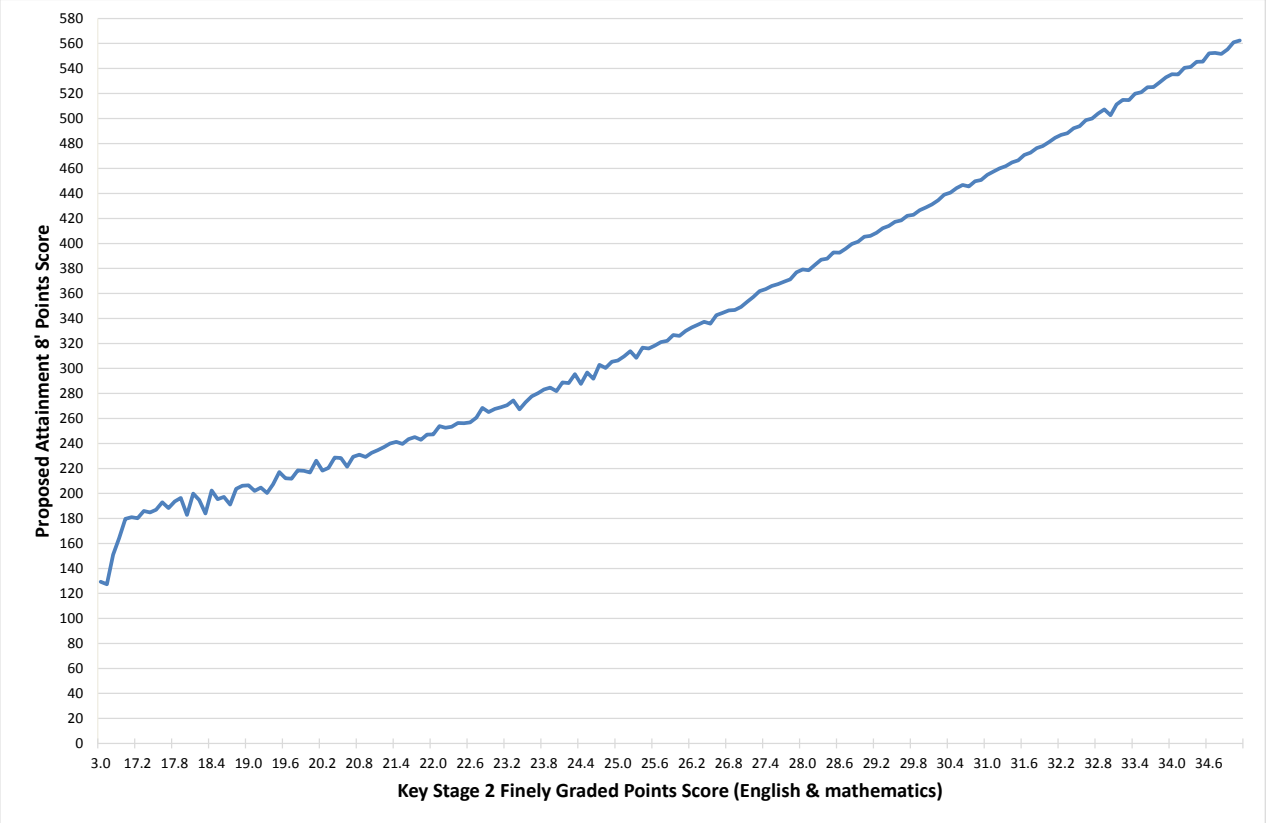
Very similar results are produced even when special schools are included (Table 4).

Table 4: Mean residuals by prior attainment quantile, simple model (all schools)

Prior Attainment Quantile	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum
				Lower Bound	Upper Bound		
20	26,353	-.10	87.33	-1.16	.95	-195.09	460.98
19	26,800	-.20	86.70	-1.23	.84	-230.46	354.96
18	26,594	-.50	87.89	-1.56	.55	-262.18	320.53
17	25,944	.27	88.91	-.81	1.35	-283.24	300.06
16	26,586	-.30	90.06	-1.39	.78	-297.08	282.92
15	27,304	.54	88.78	-.51	1.59	-322.25	270.11
14	26,717	-.78	88.50	-1.84	.29	-336.35	257.75
13	26,955	-.12	87.32	-1.16	.93	-350.44	243.65
12	26,341	.52	87.23	-.54	1.57	-367.08	229.56
11	27,113	.66	84.16	-.34	1.66	-367.08	212.92
10	26,326	-1.36	83.14	-2.37	-.36	-383.03	196.97
9	26,094	-.66	80.44	-1.63	.32	-398.28	196.97
8	26,795	-.17	78.95	-1.11	.78	-413.68	181.72
7	26,226	.02	76.29	-.90	.94	-428.36	166.32
6	27,094	.69	73.84	-.19	1.57	-445.00	151.64
5	26,367	-.03	70.44	-.88	.82	-459.41	135.00
4	25,513	.06	68.10	-.77	.90	-475.30	120.59
3	27,641	.00	63.29	-.75	.75	-491.26	104.70
2	26,324	.27	58.88	-.44	.98	-507.38	88.74
1	26,807	1.13	51.02	.52	1.74	-538.24	72.62
Total	**531,894**	**.00**	**79.84**	**-.21**	**.21**	**-538.24**	**460.98**

However, if fine graded points scores (i.e. fine grade * 6) were to replace the fine grade score, the increased level of fractal delivers a saw-toothed distribution (Figure 5), even after grouping very low values (<=17 points). For this reason we do not pursue the use of a more refined set of prior attainment pieces in this model.

Figure 5: Proposed key stage 4 'Attainment 8' points (with English and maths bonuses) by key stage 2 finely graded points score 2012 (all pupils)



The simple model has some attractive qualities. Firstly, its very simplicity is attractive and therefore it avoids the scepticism that accompanies the all too often misunderstood statistical models. Secondly, schools tend to find transition tables and charts easy to use and charts such as those shown in Figure 2 are easily produced and can be overlaid with scatterpoints (and quartiles) representing each pupil at the school.

Extending the simple model

The simple model above can easily be extended to include an English differential (the difference between key stage 2 English points and key stage 2 average points). We included the main effect plus interactions with each of our fine grade dummies.

However, the impact is rather limited, increasing marginally the amount of variance explained (in the mainstream only model) from 57.9% to 58.2%. Residual biases (Table 5) were not much improved.

Table 5: Mean residuals by prior attainment quantile, extended simple model (mainstream schools only)

Prior Attainment Quantile	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum
				Lower Bound	Upper Bound		
20	20,767	-.05	84.02	-1.20	1.09	-201.31	383.63
19	26,212	-.23	84.28	-1.25	.79	-244.27	369.08
18	26,280	-.52	85.92	-1.56	.51	-274.54	320.20
17	25,780	.42	87.23	-.65	1.48	-299.58	316.30
16	26,414	-.35	88.21	-1.41	.71	-321.92	319.87
15	27,197	.48	87.47	-.56	1.52	-333.36	263.63
14	26,646	-.88	87.40	-1.93	.17	-356.77	323.82
13	26,872	-.14	86.20	-1.17	.89	-368.45	272.18
12	26,268	.63	85.88	-.41	1.67	-372.61	273.50
11	27,074	.64	83.49	-.36	1.63	-380.10	234.17
10	26,284	-1.33	82.25	-2.32	-.34	-395.73	212.94
9	26,073	-.72	79.80	-1.69	.24	-414.38	194.11
8	26,776	-.23	78.38	-1.17	.71	-427.00	217.87
7	26,209	.07	75.68	-.84	.99	-440.46	230.12
6	27,080	.64	73.16	-.23	1.51	-457.35	170.11
5	26,360	-.09	70.03	-.93	.76	-469.80	216.44
4	25,508	.18	67.73	-.65	1.01	-477.79	238.33
3	27,638	-.02	63.06	-.76	.72	-496.25	138.57
2	26,317	.32	58.61	-.38	1.03	-510.33	166.19
1	26,804	1.13	50.94	.52	1.74	-539.87	95.35
Total	**524.559**	**.00**	**78.61**	**-.21**	**.21**	**-539.87**	**383.63**

On average, pupils achieve an extra 3 points at key stage 4 for every 1 unit increase in the English differential (i.e. when key stage 2 English finely graded points score is 2 points above the key stage 2 mathematics finely graded points score). Put another way, pupils whose key stage 2 fine grade in English is a full level (6 points) above their key stage 2 maths points score achieve an extra 9 points at key stage 4 compared to pupils with the same average fine grade.

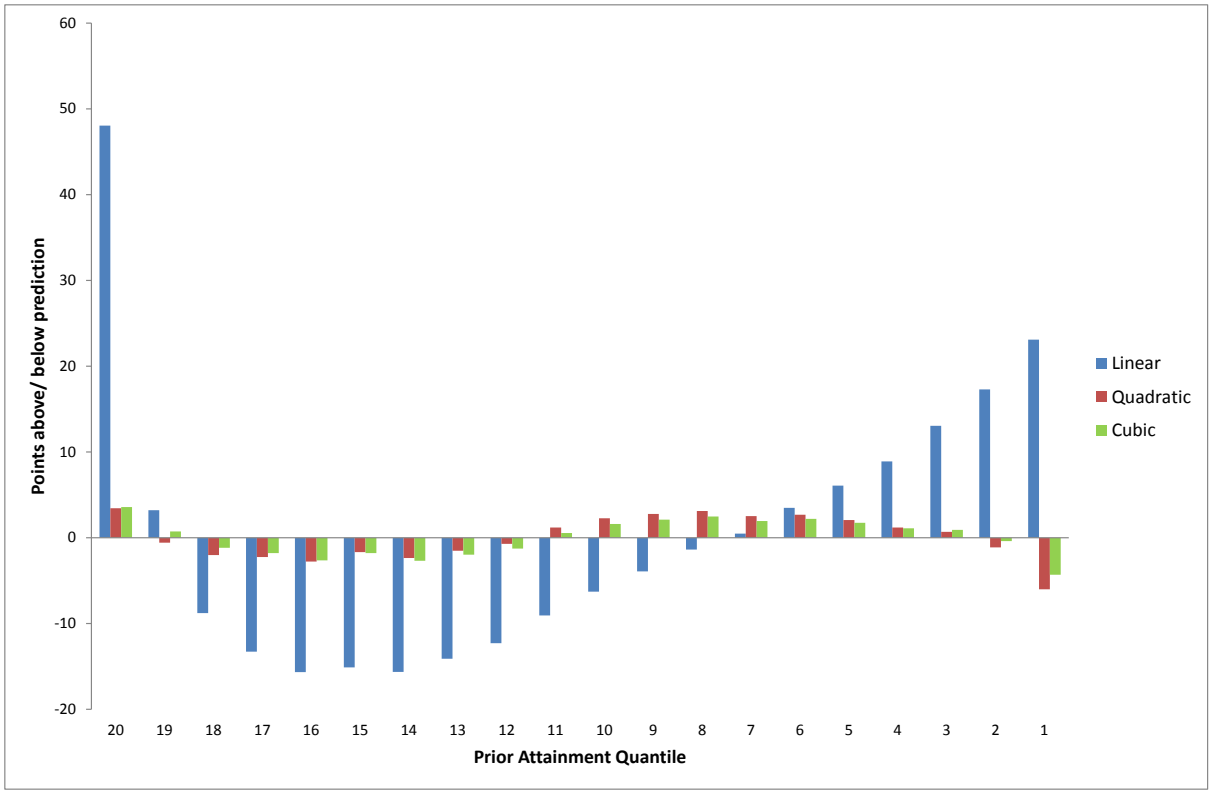
5. Moving to Ordinary Least Squares (OLS)

We now extend the simple model by regressing the proposed Attainment 8 key stage 4 measure using each pupil's finely graded average points score, their English differential, and the interaction between the two (for pupils in mainstream schools).

We note that (as shown in Table 1), the distribution of the proposed key stage 4 measure shows some slight negative skew which can be removed by the application of a power transformation (polynomials) to the independent variables. Such 'normalising' action would be expected to improve – that is, reduce – residual bias.

In Figure 6 below we show mean residual biases from this OLS regression with respect to the 20 prior attainment quantiles and how these change as polynomials in the key stage 2 fine grade points score are added to the model. Adding a quadratic term substantially improves residual bias and the further addition of a cubic term improves them slightly more.

Figure 6: Mean residuals by prior attainment quantile, OLS models (mainstream schools only)



But, even so, we calculate that even for the cubic model, residual biases are significantly different from zero for 17 of the 20 quantiles.

Standard OLS regression minimises the sums of squares of the variation in the values of observations from the mean conditional on each value having equal weight. Outlying values may, accordingly, have considerable impact on the gradient of the regression line.

Quantile regression, for example, may reduce the impact of outliers. Because the regression line is the best *average* expression of the association between the independent and explanatory variables, this may not necessarily reflect associations at extreme values and bias in estimates may result.

The effect of non–heteroscedastic errors can be reduced by using weighted least squares regression, and more complex equations may model the associations between independent and explanatory variables more appropriately at extreme values and reduce bias. But there may be occasions where the associations between the independent and explanatory variables is complex throughout their ranges and cannot be modelled with convenient mathematical shapes. Piecewise OLS regression where (usually) the same form of equation is applied separately to different parts of the explanatory variable(s) range may improve local explanation.

The biases observed in Figure 6 can be tuned out using piecewise adjustments. We introduce two dummy variables:

- Pupils in the lowest decile for prior attainment
- Pupils in the highest decile for prior attainment

We refit the cubic model with the piecewise adjustments and their interactions with each of the key stage 2 fine grade polynomials. The biases for this model (Cubic_PW) are now much smaller, as shown in Figure 7 (noting the change in the residual scale).

Figure 7: Mean residuals by prior attainment quantile, cubic OLS models (mainstream schools only)

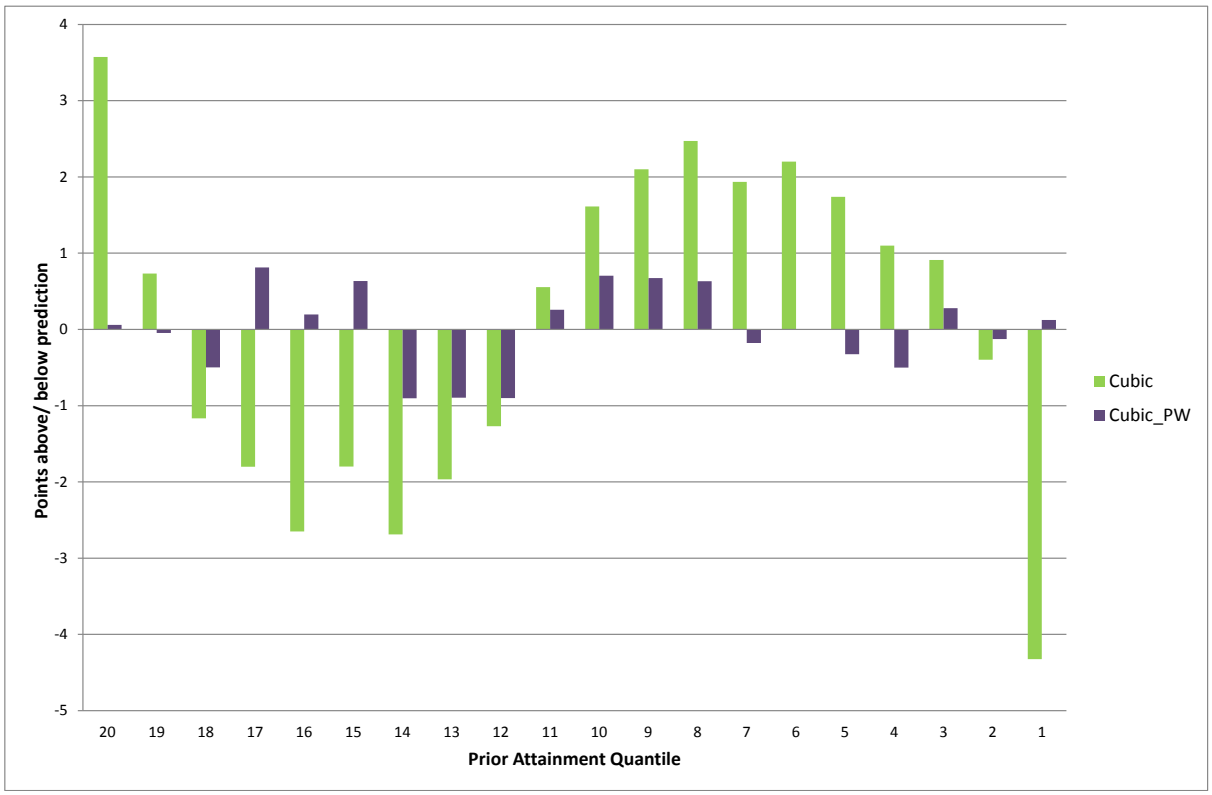


Table 6 below shows the mean residuals by key stage 2 quantile and none are now significantly different from zero and none are different from each other as indicated by a Welch test of the simultaneous equality of means.

Table 6: Mean residuals by prior attainment quantile, cubic piecewise model (mainstream schools only)

Prior Attainment Quantile	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum
				Lower Bound	Upper Bound		
20	20,767	.058	84.286	-1.088	1.204	-201.82	394.34
19	26,212	-.046	84.294	-1.066	.975	-246.75	370.90
18	26,280	-.497	85.862	-1.535	.541	-267.92	322.48
17	25,780	.814	87.219	-.251	1.878	-291.51	305.92
16	26,414	.195	88.223	-.869	1.259	-315.27	314.29
15	27,197	.635	87.404	-.404	1.674	-326.93	263.45
14	26,646	-.902	87.349	-1.951	.147	-346.07	287.39
13	26,872	-.894	86.133	-1.924	.136	-359.80	235.22
12	26,268	-.901	85.777	-1.938	.137	-366.93	251.37
11	27,074	.257	83.442	-.737	1.251	-381.59	232.29
10	26,284	.704	82.184	-.290	1.698	-395.40	218.39
9	26,073	.673	79.650	-.294	1.640	-412.54	197.54
8	26,776	.632	78.218	-.305	1.569	-422.99	228.89
7	26,209	-.178	75.524	-1.092	.737	-435.08	220.60
6	27,080	-.006	73.041	-.876	.864	-452.83	177.26
5	26,360	-.325	69.903	-1.169	.519	-465.68	223.32
4	25,508	-.501	67.575	-1.331	.328	-479.94	245.05
3	27,638	.278	62.907	-.464	1.019	-494.76	209.76
2	26,317	-.126	58.649	-.835	.582	-520.26	235.17
1	26,804	.124	51.056	-.487	.735	-539.05	228.70
Total	**524,559**	**.000**	**78.558**	**-.213**	**.213**	**-539.05**	**394.34**

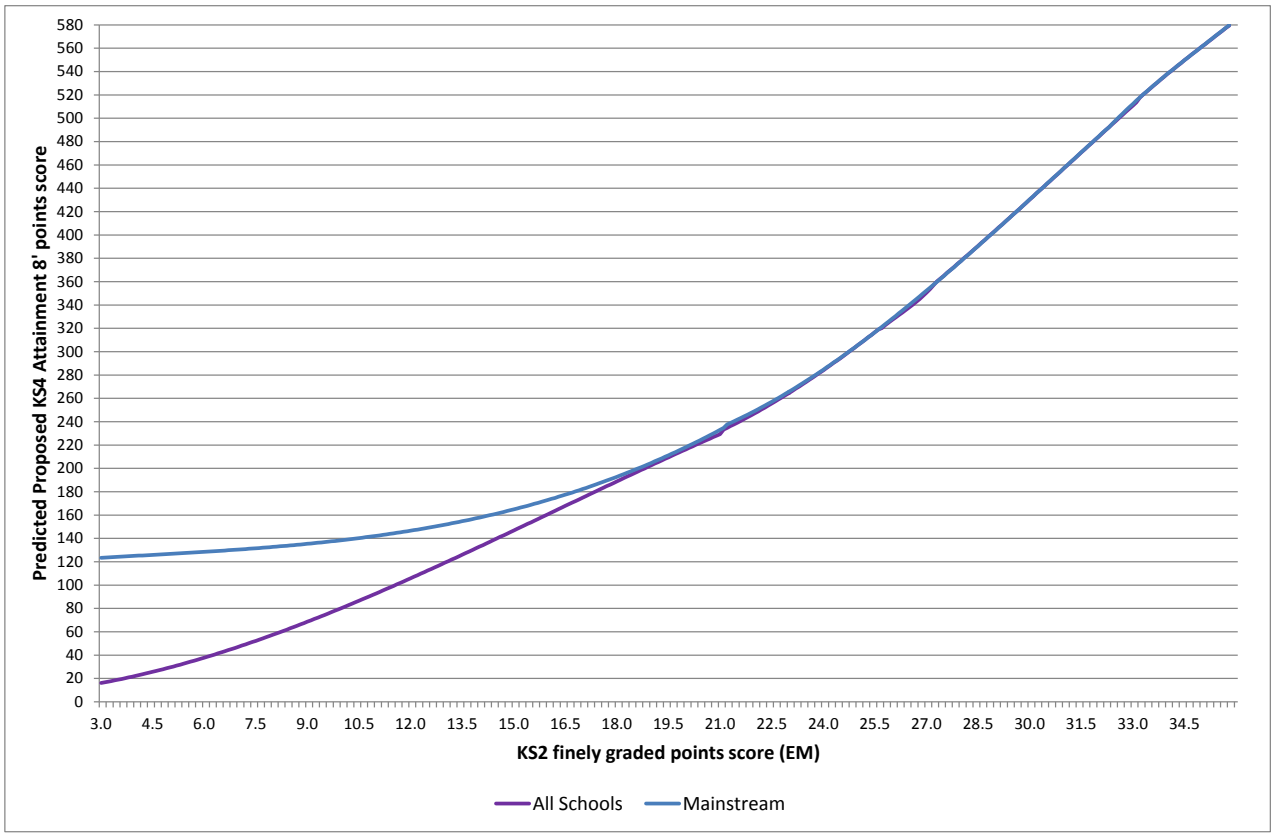
The Cubic_PW model has to be amended slightly to accommodate pupils in special schools, via a third piecewise dummy variable for pupils in the 12th to 14th quantiles. With this addition, none of the mean residuals in Table 7 are significantly different from zero. As would be anticipated from Figure 2, the standard deviation of residuals remains lower for pupils with higher levels of prior attainment.

Table 7: Mean residuals by prior attainment quantile, cubic piecewise model (all schools)

Prior Attainment Quantile	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum
				Lower Bound	Upper Bound		
20	26,353	.2521	87.07248	-.7992	1.3034	-200.60	485.87
19	26,800	-.2479	86.40915	-1.2825	.7867	-245.50	377.91
18	26,594	-.2605	87.47394	-1.3119	.7908	-267.64	327.65
17	25,944	.6756	88.21239	-.3979	1.7490	-292.04	309.77
16	26,586	-.5808	89.35036	-1.6549	.4933	-316.58	317.09
15	27,304	-.0410	88.19198	-1.0872	1.0051	-327.71	263.69
14	26,717	-.3410	87.96847	-1.3959	.7139	-345.34	293.46
13	26,955	.6302	86.72198	-.4052	1.6655	-358.06	238.19
12	26,341	-.2990	86.58891	-1.3447	.7467	-366.56	256.95
11	27,113	-.1609	83.75952	-1.1579	.8361	-382.00	235.31
10	26,326	.2271	82.73194	-.7723	1.2265	-395.75	218.60
9	26,094	.4013	79.92818	-.5686	1.3711	-412.82	197.71
8	26,795	.4194	78.49698	-.5206	1.3593	-423.13	229.77
7	26,226	-.3414	75.87512	-1.2597	.5770	-435.05	220.79
6	27,094	-.0707	73.33446	-.9439	.8026	-452.68	176.82
5	26,367	-.2487	69.96923	-1.0933	.5959	-465.42	222.14
4	25,513	-.4203	67.70667	-1.2511	.4106	-479.58	243.08
3	27,641	.3902	62.95733	-.3520	1.1324	-494.43	206.57
2	26,324	-.1292	58.84469	-.8401	.5817	-519.87	231.44
1	26,807	.1269	51.07909	-.4846	.7384	-539.12	223.55
Total	**531,894**	**.0000**	**79.38990**	**-.2134**	**.2134**	**-539.12**	**485.87**

Figure 8 below shows the predicted points scores from the cubic piecewise model for mainstream schools and all schools separately where English subject differentials have been held constant at zero. As can be seen, the inclusion of the pieces does not lead to any marked increases in predicted scores where the pieces have been fitted.

Figure 8: Predicted scores from the cubic piecewise models



We have shown that all pupils (with key stage 2 results) can be included in a Cubic_PW model which produces unbiased mean residuals. But, as Figure 8 shows, pupils in special schools tend to attain on average lower key stage 4 outcomes than pupils of similar prior attainment in maintained schools: indeed, on average, pupils attending special schools achieved 85 points below prediction.

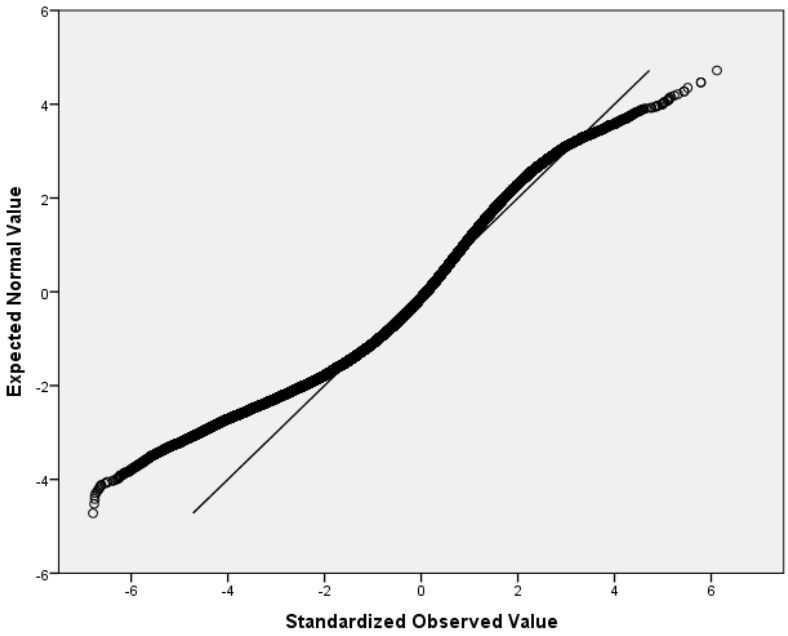
The implication of the Cubic_PW model as applied above would be that residuals for pupils in special schools would tend be substantially below average, and thus the vast majority of such schools would have 'negative' value-added scores and be said to be performing poorly. Such bias – if bias it is considered to be - could be removed by including a dummy variable for special schools (and tuning the Cubic_PW model). Whichever direction is taken must at heart be a policy rather than a statistical decision.

Figure 9 below plots the distribution of standardised residuals from the Cubic_PW model. Most residuals are judged normal having values between -2 and 2, but those at the extremes exhibit high levels of non-normality particularly at the extreme lower end of key stage 2 attainment where pupils' value-added scores are lower than we would expect from a normal distribution.

This is not a feature endemic to this particular model: any model will have to deal with pupils whose key stage 4 outcome is radically different from expectation (for example, able pupils scoring 0 points). Such pupils will exert a disproportionate influence on

schools' value-added scores. Nevertheless, a reasoned rescaling of the GCSE grade scale generally reduces non-constant variance.

Figure 9: Q-Q Plot of standardised residuals, cubic piecewise model (all schools)



Finally, the minor biases shown in Table 7 can be eliminated completely by extending the piecewise process to include each of the 20 percentile bands in an OLS regression model (Table 8). In the Percentile model, we include:

- key stage 2 finely graded average point score (average point score) (linear term only)
- English subject differential
- Percentile band
- Interaction between key stage 2 average point score and English subject differential
- Interaction between key stage 2 average point score and percentile band.

Table 8: Mean residuals by prior attainment quantile, Percentile Model (all schools)

Prior Attainment Quantile	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum
				Lower Bound	Upper Bound		
20	26,353	.00	87.22	-1.05	1.05	-196.46	497.10
19	26,800	.00	86.40	-1.03	1.03	-247.16	378.61
18	26,594	.00	87.47	-1.05	1.05	-266.67	327.72
17	25,944	.00	88.21	-1.07	1.07	-291.45	308.32
16	26,586	.00	89.35	-1.07	1.07	-316.53	317.47
15	27,304	.00	88.19	-1.05	1.05	-327.46	263.73
14	26,717	.00	87.97	-1.05	1.05	-344.41	294.29
13	26,955	.00	86.72	-1.04	1.04	-358.42	237.62
12	26,341	.00	86.59	-1.05	1.05	-366.71	257.16
11	27,113	.00	83.75	-1.00	1.00	-380.69	234.48
10	26,326	.00	82.73	-1.00	1.00	-395.57	217.89
9	26,094	.00	79.93	-.97	.97	-413.07	197.23
8	26,795	.00	78.50	-.94	.94	-423.47	228.99
7	26,226	.00	75.87	-.92	.92	-434.15	221.01
6	27,094	.00	73.33	-.87	.87	-453.83	175.64
5	26,367	.00	69.96	-.84	.84	-464.64	222.23
4	25,513	.00	67.71	-.83	.83	-478.98	242.87
3	27,641	.00	62.95	-.74	.74	-495.36	204.87
2	26,324	.00	58.84	-.71	.71	-519.31	230.34
1	26,807	.00	51.07	-.61	.61	-538.96	222.11
Total	**531,894**	**.00**	**79.40**	**-.21**	**.21**	**-538.96**	**497.10**

Both the cubic piecewise and percentile models deliver the desirable minor improvements to residuals over those from the simple model. In the percentile model there is no average residual bias in any of the key stage 2 prior attainment quantiles.

Significant enhancements to the properties of residuals usually come at a cost on model complexity or complication. We certainly accept that the percentile model is more difficult to understand than a simple regression but we think that, for those stakeholders with an interest, the explanation of this model can be made straightforward.

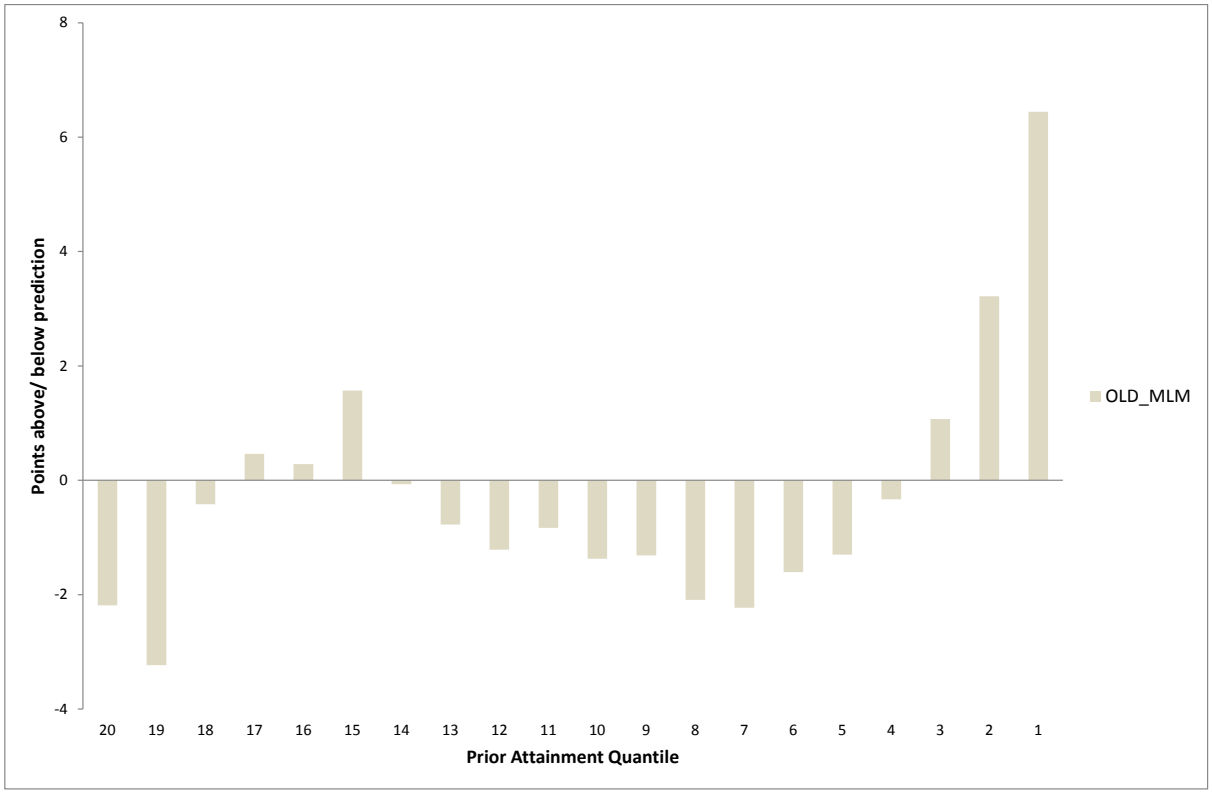
6. Multilevel Modelling (MLM)

The current value-added model is a random intercept multilevel model with the following independent variables:

- Finely graded key stage 2 average point scores (with quadratic and cubic terms)
- English differential (difference between finely graded key stage 2 points in English and key stage 2 average point scores)
- Maths differential (difference between finely graded key stage 2 points in maths and key stage 2 average point scores)

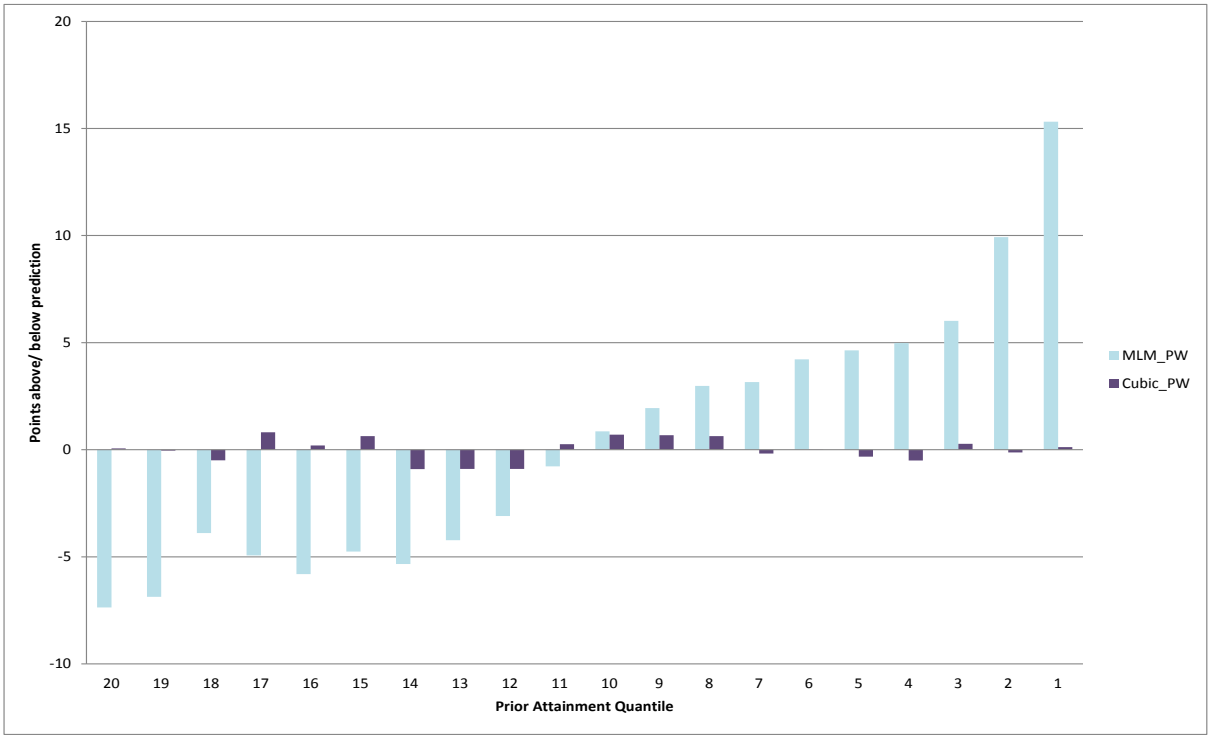
Figure 10 below shows the mean residuals by quantile for the current model. There are non-trivial biases at the extremes of key stage 2 attainment. Mean residuals are significantly different from zero for 14 of the 20 quantiles. We observe that these residuals are from the *fixed* part of the MLM model and bias in them arises from the clustering of pupils by key stage 2 attainment within schools.

Figure 10: Mean key stage 2-key stage 4 Residuals 2012 (current points score) by Prior Attainment Quantile (mainstream schools only)



Fitting the Cubic_PW model (mainstream schools only) as a random-intercept model (MLM_PW) – ignoring more complex differential effectiveness MLM models – continues to result in significant residual bias, as Figure 11 below indicates.

Figure 11: Mean key stage 2-key stage 4 Residuals 2012 (proposed points score) by Prior Attainment Quantile (mainstream schools only)



One of the principal attractions of MLM[1] is the creation of individual school lines within a standardised model framework. This improves the use of information from the data and, in principle, can yield more robust ('shrunken') value-added scores and standard errors.

If shrunken estimates are seen as adding value to the interpretation of value-added scores, these can be achieved by other means through Empirical-Bayes adjustments[2] to any of the other models we have outlined.

[1] Given that we are not concerned with the parameter estimates (or their standard errors) from the models

[2] See Greenland S, Robins J.M., (1991). Empirical-Bayes adjustments for multiple comparisons are sometimes useful. Epidemiology, 2:244-251.

7. Lowess

Non-parametric regression models ('local regression estimator methods') are an alternative method of constructing an average relationship across the prior attainment range. These models make no formal assumptions regarding the processes that generated the data on which they are applied. They provide a basis for manufacturing a linear – or, more usually, a polynomial - association between an independent and explanatory variable using averaging and smoothing methods using data values alone. These models consider a slice (bandwidth) of the range of the explanatory variable and use it to estimate an *average* value of the independent variable for each value of the explanatory variable within it by weighting, summing, and averaging the values of the independent variable for its neighbours. A curve across *all* data points is developed by joining up the independent value *estimates* across the whole range of the explanatory variable.

Because non-parametric methods are local averaging methods, sizes of different groups of neighbouring points (bandwidth) lead to different estimates for each data point. The choice of bandwidth is generally considered more important than choice of how the neighbouring points are averaged. A smaller bandwidth improves accuracy – that is, reduces bias – because only close neighbours are used to give estimates. But this is at the cost of increased variation in the curve (less smoothness). A larger bandwidth improves the smoothness of the curve but increases bias because points which have greater distance from the data point under consideration are used in its estimation. Models from traditional regression methods are tested for bias and efficiency (confidence intervals), as are those from non-parametric models. A natural metric to determine the range of the most efficient slice (bandwidth) is the value of mean-squared error (MSE) - the sum of bias squared and variance- taken over all point estimates.

Lowess carries out a separate locally weighted regression on *each* of the 530,000 pupils in the national dataset. The weights reflect the distance each neighbour is from the specific value of the explanatory variable applied to an estimate of the neighbour's independent variable resulting from a polynomial least squares regression conducted across all the data points in the bandwidth. A 'bandwidth' of 0.035 was chosen so that predictions continuously increased with respect to prior attainment (lower bandwidths did not meet this criterion). This means that in each local regression 3.5% of cases in the dataset are used with those cases closest to a given pupil (in prior attainment) given greatest weighting.

Figure 12: Predicted outcomes, lowess model (all schools)

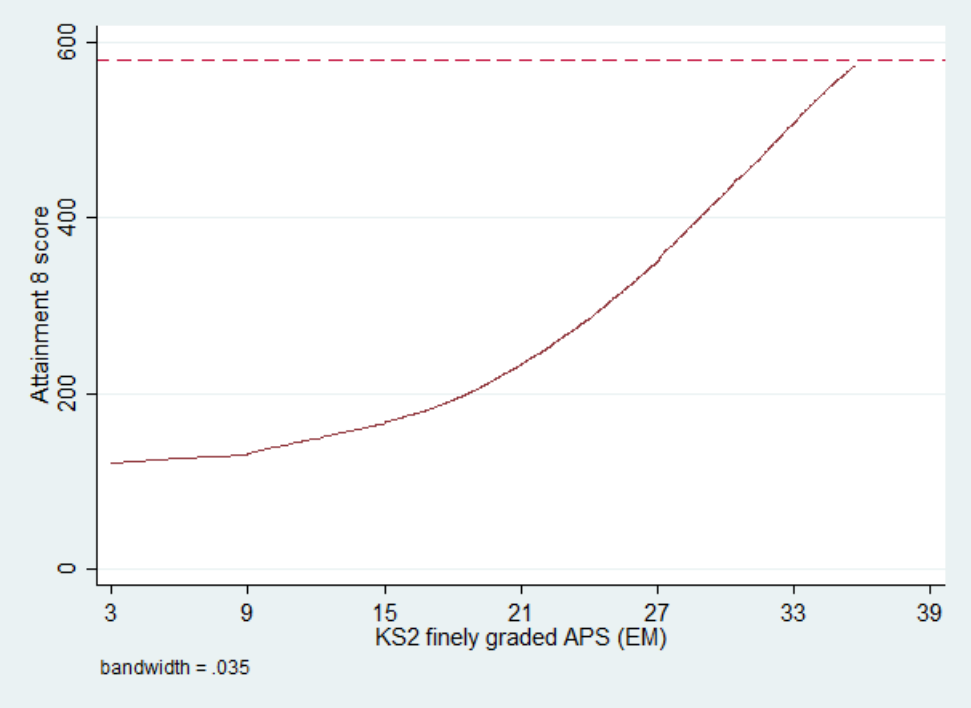


Table 9 below shows that mean residuals by prior attainment quantile are not significantly different from zero.

Table 9: Mean residuals by prior attainment quantile, lowess model (all schools)

Prior Attainment Quantile	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum
				Lower Bound	Upper Bound		
20	26,353	-0.45	87.02	-1.50	0.60	-193.60	487.24
19	26,800	-0.25	86.70	-1.29	0.79	-232.60	353.67
18	26,594	-0.13	87.82	-1.19	0.92	-260.32	316.31
17	25,944	0.16	88.85	-0.92	1.24	-280.99	297.93
16	26,586	-0.06	90.04	-1.14	1.02	-300.38	287.00
15	27,304	0.09	88.70	-0.96	1.15	-317.31	265.06
14	26,717	-0.18	88.44	-1.24	0.88	-331.97	255.23
13	26,955	0.11	87.22	-0.93	1.15	-345.31	243.68
12	26,341	0.02	87.07	-1.03	1.08	-361.73	234.22
11	27,113	-0.24	84.11	-1.24	0.76	-374.50	211.59
10	26,326	0.17	83.08	-0.83	1.17	-388.06	204.57
9	26,094	0.07	80.29	-0.90	1.05	-401.50	189.72
8	26,795	0.06	78.80	-0.89	1.00	-414.45	176.96
7	26,226	-0.06	76.14	-0.98	0.87	-426.55	164.82
6	27,094	0.04	73.72	-0.84	0.91	-441.87	152.82

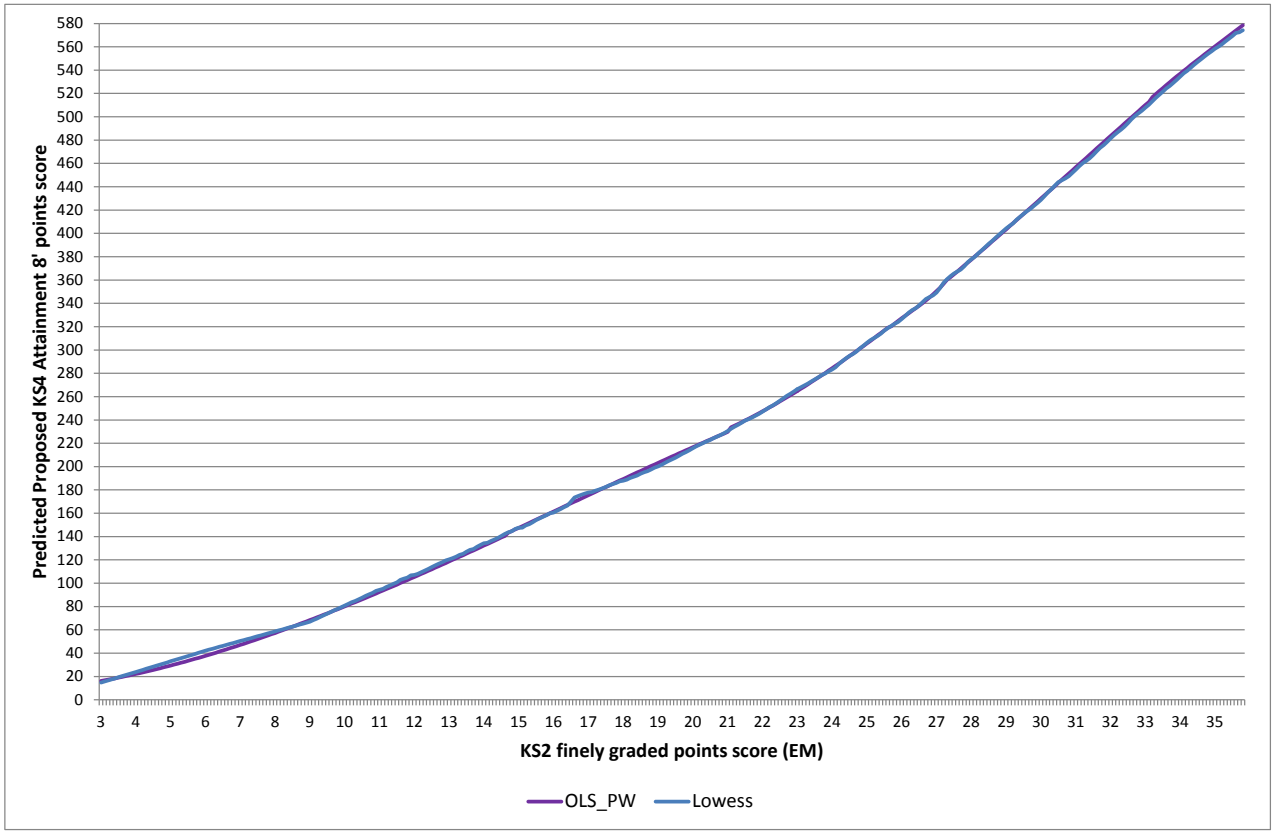
5	26,367	-0.04	70.33	-0.89	0.81	-455.42	136.56
4	25,513	0.03	67.95	-0.80	0.87	-469.83	123.59
3	27,641	0.10	63.13	-0.64	0.85	-487.80	108.35
2	26,324	-0.28	58.73	-0.99	0.43	-508.57	89.99
1	26,807	0.19	50.78	-0.41	0.80	-540.95	68.22
Total	**531,894**	**-0.03**	**79.72**	**-0.25**	**0.18**	**-540.95**	**487.24**

We experimented with larger bandwidths but these increased processing time (typically from 15 minutes to 40 minutes) and led to less satisfactory residual biases.

A limitation of lowess is that only a *single* independent variable (in this case key stage 2 finely graded average point score in English & maths) can be used and that model parameters are not output by the statistical software.

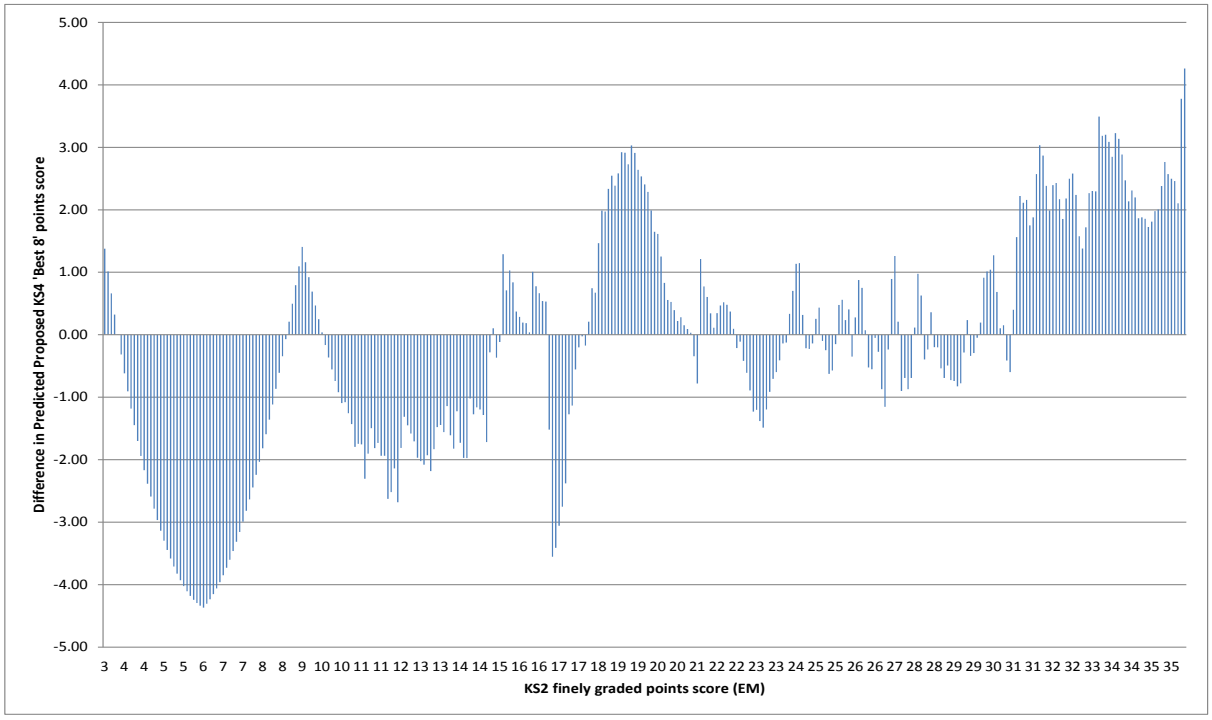
However, mean predictions for each mean key stage 2 finely graded points score can easily be produced and Figure 13 below compares these values with predictions from the cubic piecewise model.

Figure 13: Comparison of predicted values from the lowess and cubic piecewise models



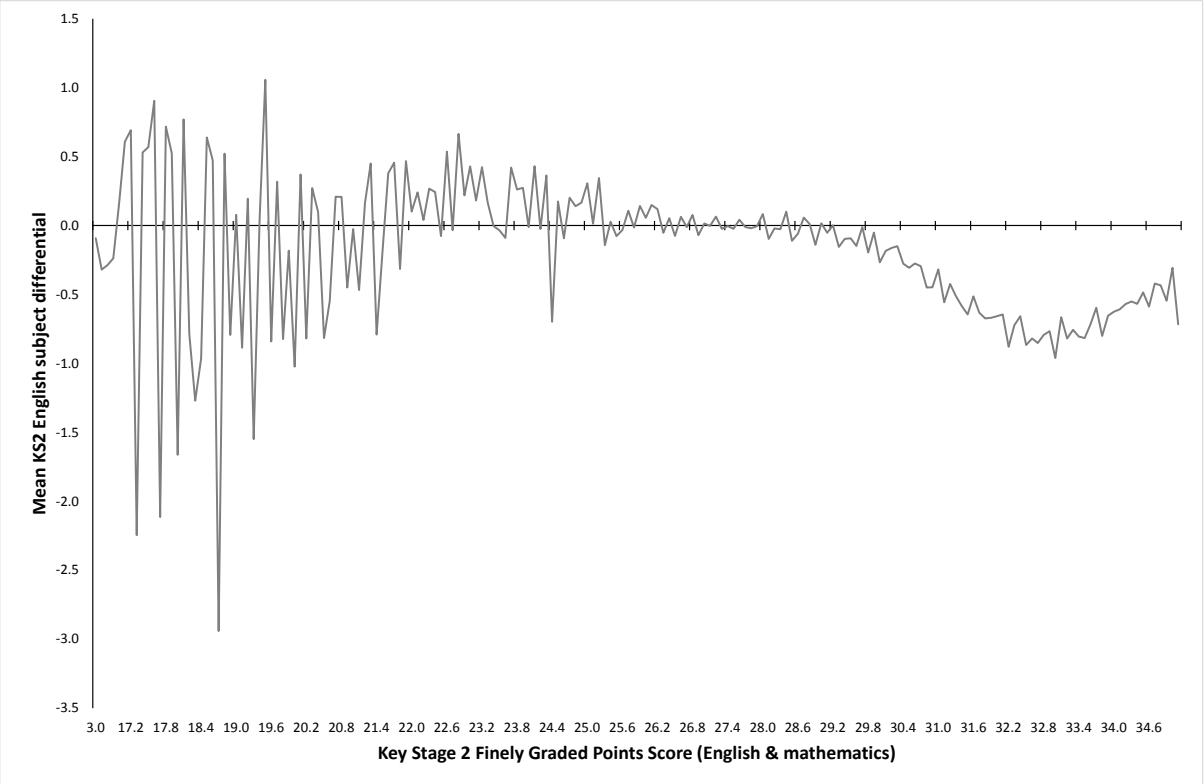
Differences in predictions between cubic piecewise and lowess are relatively small. The cubic piecewise model produces slightly higher predictions for pupils with high prior attainment while the inverse is true for pupils with low prior attainment, as Figure 14 indicates.

Figure 14: Differences in predicted values from the lowess and cubic piecewise models



At the upper end, the differences appear to be caused by the English subject differential. In Figure 15, we hold this constant at zero. However, pupils with high prior attainment (level 5+; >=30 points) tend to achieve lower fine grades in English than in mathematics. This artificially inflates the predictions shown in this figure. For example, the assumption of zero difference adds a further 4.9 points to the prediction for a pupil with a key stage 2 average point score of 33.

Figure 15: mean English subject differential by key stage 2 prior attainment



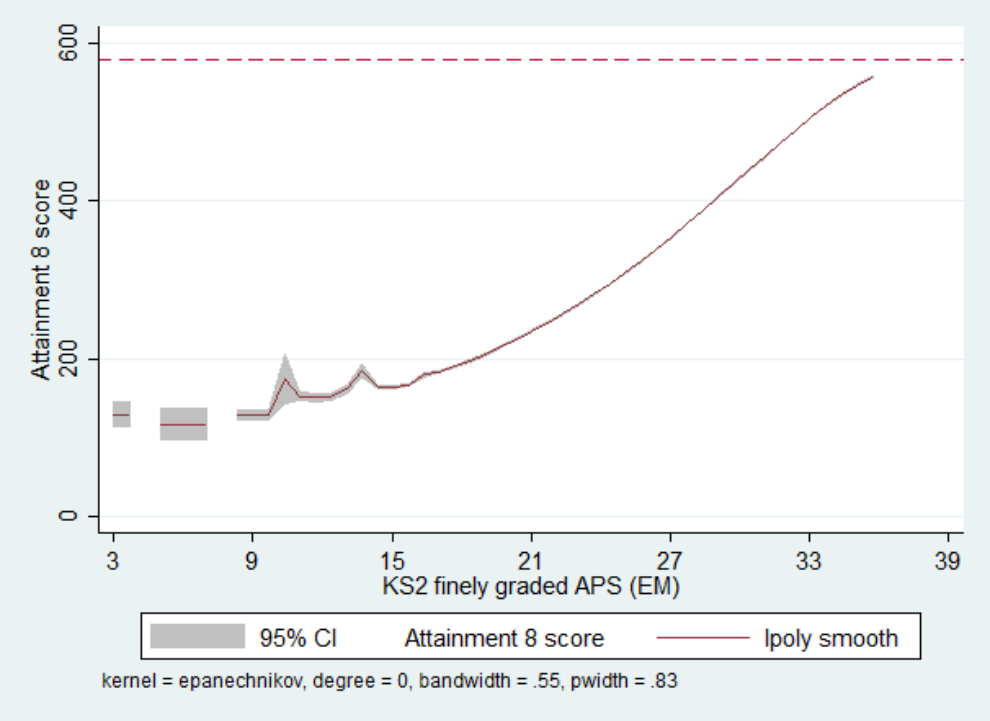
8. Kernel Regression

Kernel regression is a special case of lowess where the weights applied to the values of the independent variable just reflect the distance between the neighbouring explanatory data points. *In extremis*, Lowess is more attractive than kernel regression in that it is more robust to outlying data points and minimizes boundary problems at the extremes of the explanatory variable. Kernel regression also has similarities to OLS regression but instead of each observation having equal weight as in ordinary least squares (OLS), or weights proportional to the inverse of variance - as is often the case in weighted least squares (WLS) - a different rationale determines the choice of weights.

In piecewise OLS regression, each data point within the explanatory variable slice continues to have equal weight, and a linear line (or polynomial curve) is constructed over the slice. The non-parametric techniques, however, construct an estimated value of the independent variable for each value of the explanatory variable by weighting the independent variable values of neighbouring values of the explanatory variable using a probability density function (a Kernel). This requires that weights decrease as the distance of neighbours from the data point increases. The kernel function is defined to be continuous, symmetric (around zero), integrates (to unity), and satisfies additional boundedness conditions. Epanechnikov, Gaussian, quartic and uniform are standard Kernels. There is usually little difference in estimates derived from alternative Kernel functions, with the Epanechnikov distribution accepted as optimal although we find empirically using the key stage 2-4 dataset that we can use a smaller bandwidth if we use a Gaussian kernel.

We use the proposed ‘Attainment 8’ measure as the outcome and key stage 2 fine average point scores in English and maths as the independent variable. We begin by letting STATA fit the relationship using its defaults (see pp931-933 of the STATA user guide)

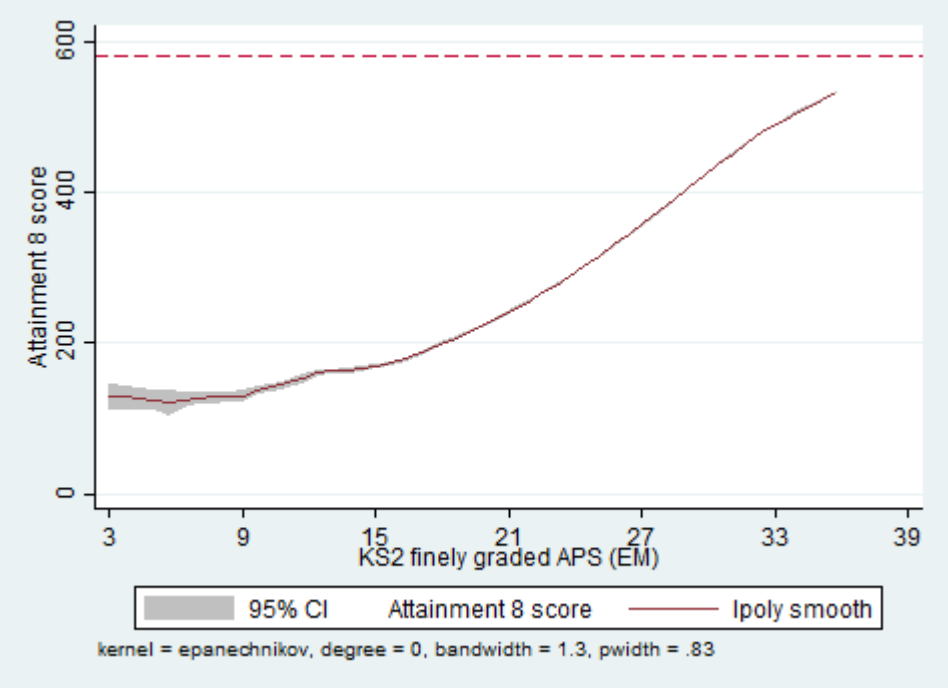
Figure 16: Kernel regression with standard defaults (mainstream schools only)



Note “degree=0” denotes local-mean smoothing. The bandwidth of 0.55 is determined by a “rule of thumb” (ROT) test of the form shown in equation 3 on p. 939. It is the “bandwidth that minimises the conditional weighted mean integrated square error.”

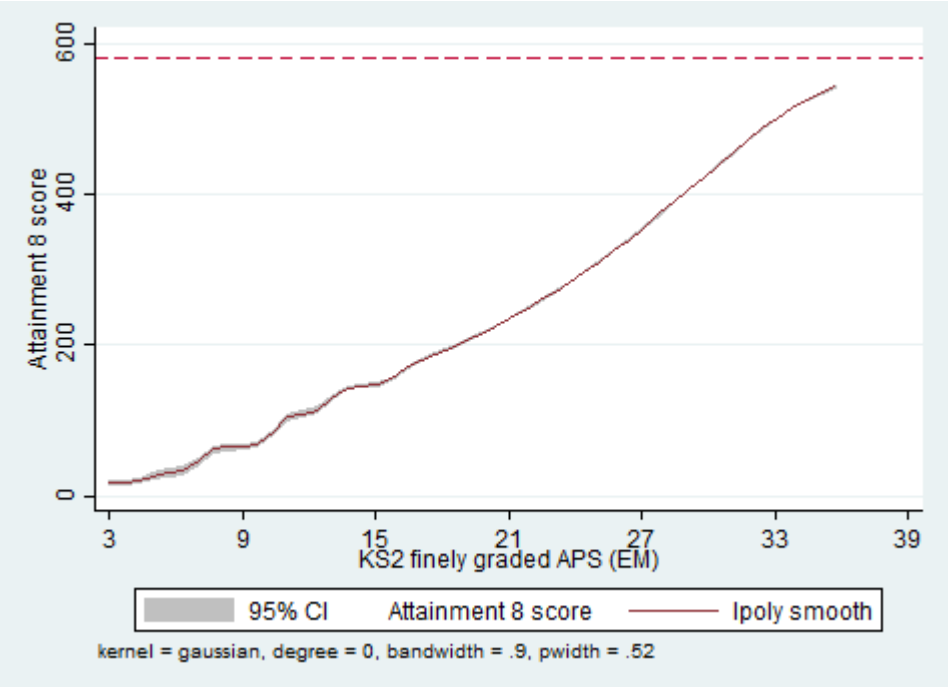
The line can be smoothed further by increasing the bandwidth again. In the example below, Figure 17, we have increased the bandwidth to 1.3. This also narrows the confidence intervals as more observations are used in each local regression. This resolves the problem with slope gradient other than for pupils with the very lowest levels of prior attainment.

Figure 17: Kernel regression with bandwidth 1.3 (mainstream schools only)



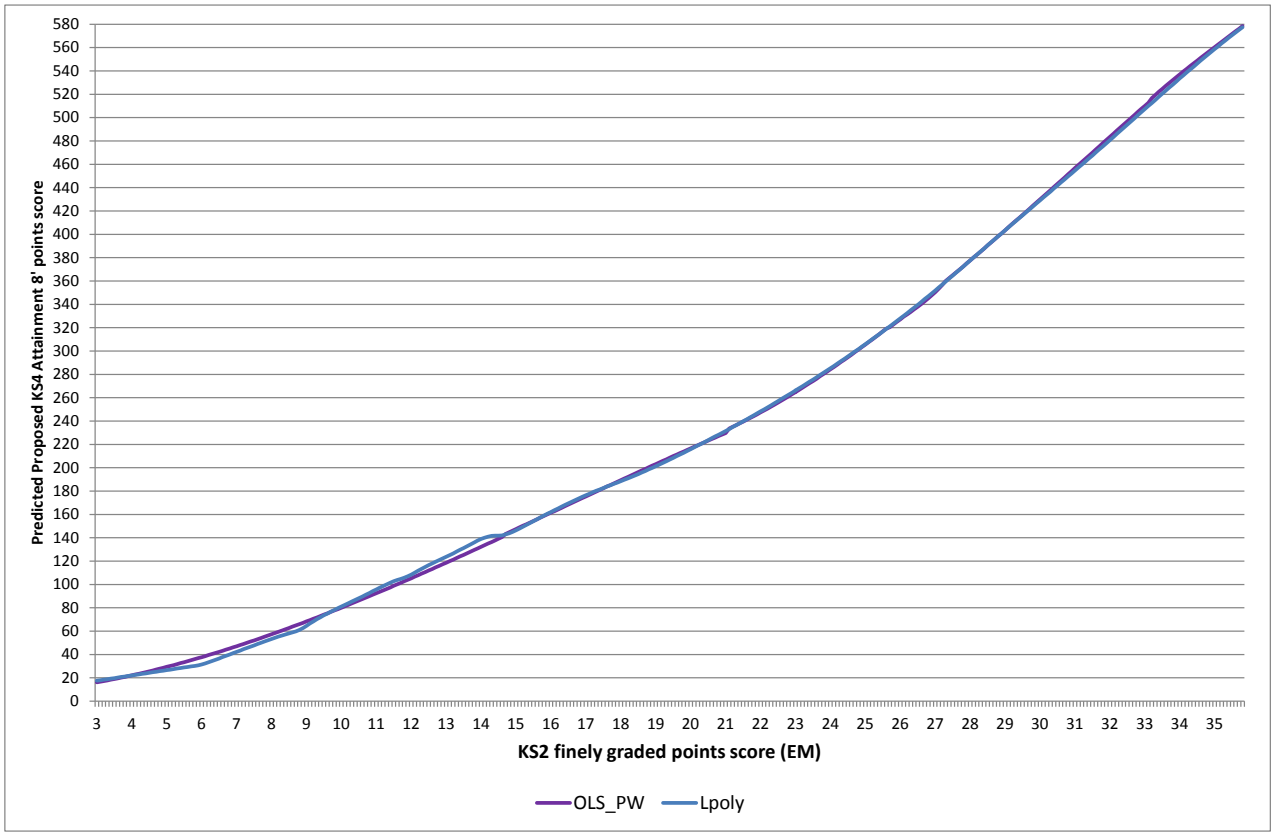
This problem disappears, and a narrower bandwidth can be used, if special schools are included and a Gaussian kernel (rather than Epanechnikov) is used. As shown in the section on OLS regression, the lower bound of the range of predictions is lower when pupils in special schools are included.

Figure 18: Kernel regression with Gaussian kernel (all schools)



The above line in Figure 18 is very similar, with some slight variation at the lower end of prior attainment, to the OLS Cubic_PW line, as is shown in Figure 19 below.

Figure 19: Comparison of Kernel regression and Cubic_PW line



The above analyses show that kernel regression can produce an acceptable pupil progress relationship. We have not fine-tuned this regression as we have for the OLS or lowess models although this could easily be achieved. Kernel regression does not appear to yield any significant benefit over the lowess model. Moreover, the choice of bandwidth (0.9 points) relates to the scale of the independent variable which, in this case, is not truly continuous. A difference of 0.9 points either side of any given value of key stage 2 average point score could cover a wider range of ability than for a value at a different part of the range.

9. Quantile Regression

Quantile regression (QR) can be used to estimate a distribution function for every pupil on the basis of their prior attainment. QR can estimate any percentile of key stage 4 outcome conditional on key stage 2 attainment – including the median (or 50th percentile/quantile) – to provide a 'distribution' of performance and thus progress. This method underpins the Colorado Student Growth Model[3]. Outcomes are expressed as 'growth percentiles': one of 80[4], for instance, indicates that a pupil's actual key stage 4 score was higher than 80% of pupils with identical prior attainment.

QR is a suitable method for the analysis of heteroscedastic data as it does not make any assumptions about the distribution of regression errors, or residuals[5]. The use of relative position methods - median regression - is less sensitive to outliers than OLS (though we note from Figure 2 that the choice of key stage 4 outcome measure and the normalisation of key stage 2 and key stage 4 measures generally will reduce their importance).

But, and as we outline below, differences between percentiles do continue to be a feature of the data - narrower at the upper end of the prior attainment distribution than the lower end. The treatment (or not) of the smaller span of residuals – or at least its full understanding and implications – remains an important consequence of the value-added modelling.

In principle, we could calculate for every pupil a growth percentile using simultaneous quantile regression (sqreg) in STATA. While feasible, such models are rather intensive computationally.

For ease of illustration we run a simultaneous quantile regression model on pupils in *all* schools to estimate the lower quartile (quantile 25), the median (quantile 50) and upper quartile (quantile 75) based on the Cubic_PW model developed previously. The proposed 'Attainment 8' points score is the dependent variable and the following are the regressors:

- Finely graded mean key stage 2 average point scores (with a quadratic term);
- English differential (difference between finely graded key stage 2 points in English and key stage 2 average point scores);
- Interaction between key stage 2 average point scores and English differential;
- Piecewise adjustments for the top 10% and bottom 10% of pupils based on key stage 2 average point scores.

---

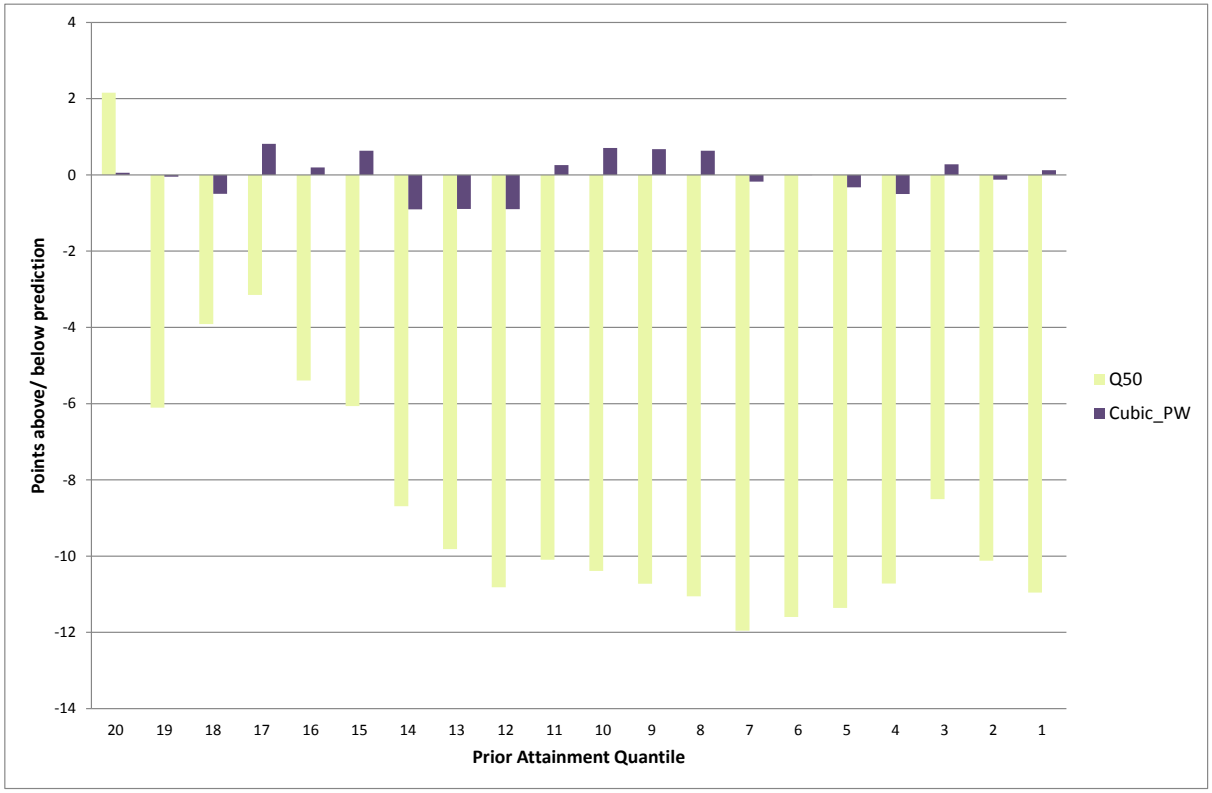
[3] http://www.cde.state.co.us/research/download/pdf/Aprimeronstudentgrowthpercentiles_a.pdf

[4] As we use STATA to fit QR models we shall adopt the STATA convention that higher percentiles are better, i.e. the top 1% of pupils achieve results above the 99th percentile.

[5] Cameron, A.C. and Trivedi, P.K. (2010) Microeconometrics using Stata, Texas: Stata Press

We do not claim that this choice is optimal - rather we present some illustrative findings from a reasonably well-fitting model that would benefit from further tuning. Neither do we propose that the median line from a quantile regression model is used to create pupil (or school) predictions. As indicated by Figure 2, the median line tends to be slightly higher than the mean line, creating the biased residuals as shown in Fgiure 20 below.

Figure 20: Mean key stage 2-key stage 4 Residuals 2012 (proposed points score) by Prior Attainment Quantile (all schools)

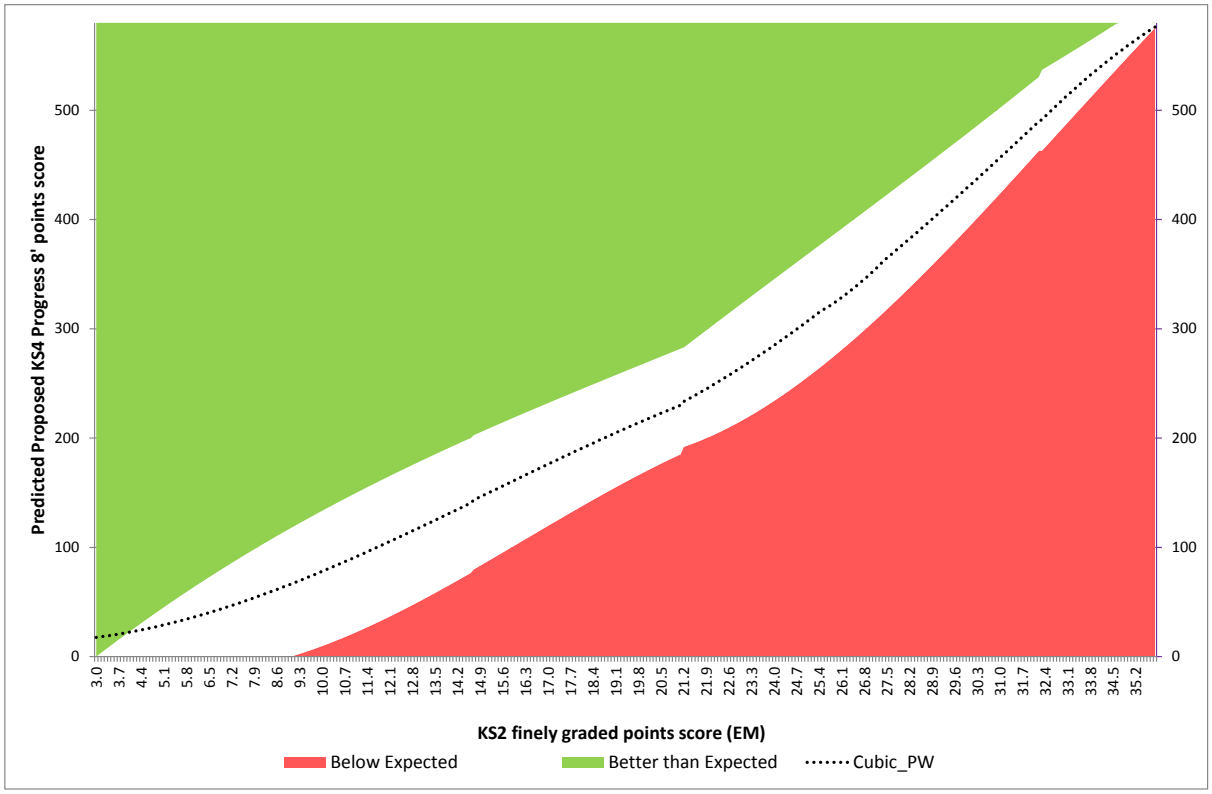


But, QR could be used to develop *threshold* measures of pupil progress - for example, the proportion of pupils making better than expected progress. A specific quantile – say the upper quartile - could be used to represents 'better than expected' progress, and each school would have a proportion of pupils who achieve this threshold.

Such a basis could be combined, for example, with a mirror-image threshold for the proportion of pupils making 'less than expected progress'. This might introduce to schools easy-to-understand levels of challenge which would be fair irrespective of the key stage 2 attainment of their pupils.

To illustrate this concept, for all pupils, we show in Figure 21 below the definition of better than expected progress (UQ) and below expected progress (LQ) for all pupils from a QR regression combined with the predicted (mean) outcome line from the Cubic_PW model.

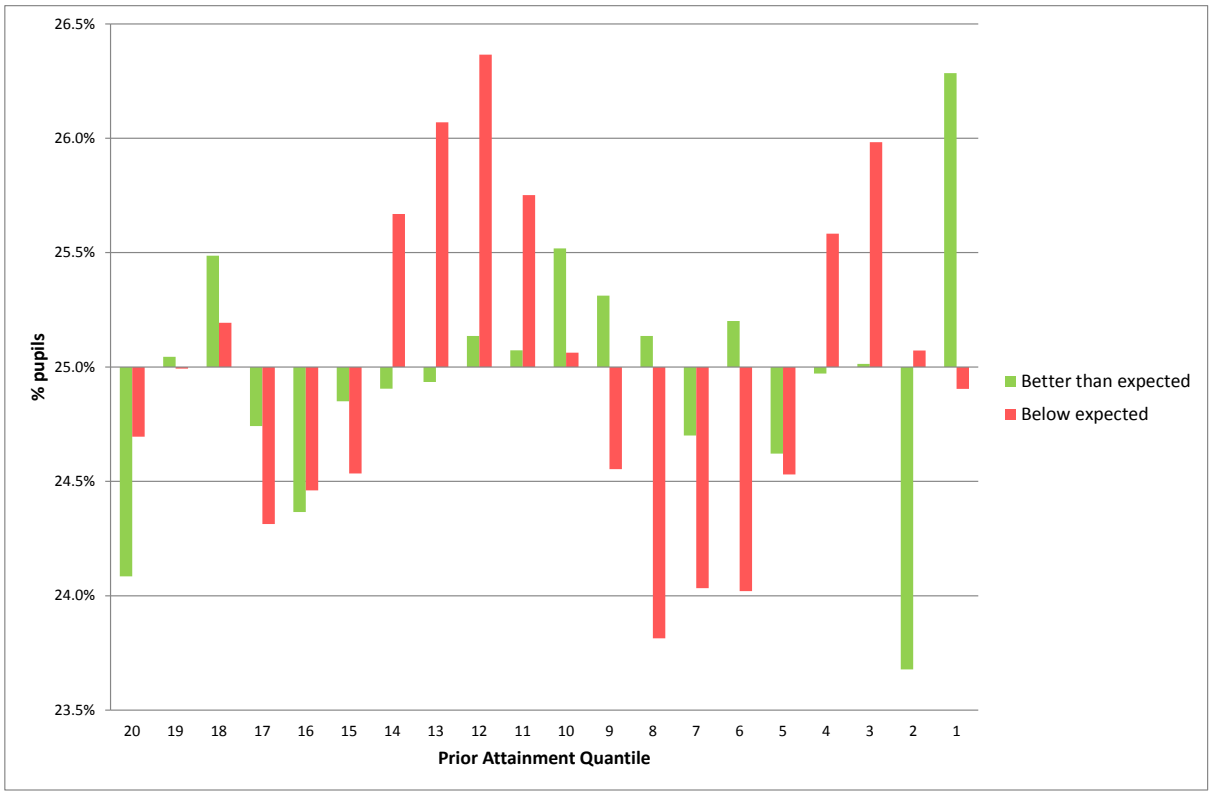
Figure 21: Expected Progress Chart, Cubic Piecewise Model (All Schools)



We observe that the green ‘better than expected zone’ ends at key stage 2 average point score 34.6. Some 0.4% of pupils in the 2012 key stage 4 cohort had a higher key stage 2 average point score. Analogously, the zone includes 0 points for pupils with a key stage 2 average point score of 3.0. Some post hoc rules could be implemented such that pupils who achieve 8 A* grades (580) points are all deemed to have made ‘better than expected progress’, and that pupils who have achieved 0 points cannot be determined to have made ‘better than expected progress’.

Fgiure 22 shows the proportion of pupils in the green and red zones of Figure 15 by prior attainment quantile. We would expect that the proportion of pupils in each quartile to be close to 25%. Many but by no means all are tolerably close to this proportion (in the range 24.5% to 25.5%). This range could be reduced with further tuning to the model.

Figure 22: Proportion of pupils making below/ better than expected progress (all schools)



10. Non-constant variance

The variance function indicated by the standard deviation columns of Tables 6 and 7 will have an impact on tests of statistical significance on value-added scores for schools with disproportionate numbers of pupils at the lower or upper ends of the prior attainment range.

Conventionally, assuming homoscedasticity, the standard error for a school's value-added score would be the standard deviation in national pupil residuals divided by the square root of the school's key stage 4 value-added pupil numbers.

Standard errors for schools with disproportionately large numbers of pupils with low prior attainment (where residual variance is highest) would be too small, and conversely for schools with high proportions of pupils with high key stage 2 attainment (where residual variance is lowest), too large. Standard errors may also be affected by the extreme residual values shown in Figure 9.

This may or may not be an issue- and it depends on whether the issue of 'fairness' relates to pupils or schools. We would need to establish whether the variance is a function of:

- Pupils themselves; or
- The schools they attend.

If it is the case that part of the increased variance is due to differential school effectiveness, it might be acceptable, from a policy perspective, to ignore the non-constant variance. We note that non-constant variance persists even when we control for additional pupil and school level contextual factors.

If non-constant variance was considered a problem, it could be attenuated by one or more of the following:

- Additional explanatory variables;
- Rescaling the outcome variable (i.e. the points scores associated with grades);
- *Post hoc* adjustments, e.g. calculating standard errors (and thus significance tests) standardised for pupils in different prior attainment bands; or
- A different approach to modelling heteroscedastic distributions (e.g. quantile regression).

11. Model Comparisons

In Table 10, we compare two key statistics- the mean squared residual and the proportion of variance explained- from the various models described above.

Table 10: Model Diagnostics (mainstream schools only)

	% variance explained	Root mean square residual	Unbiased residuals by prior attainment
Simple	57.9%	79.0	Yes
Simple Extended	58.2%	78.6	Yes
OLS (cubic piecewise)	58.3%	78.6	Yes
OLS (percentile)	58.3%	78.6	Yes
Kernel	58.0%	78.9	Yes
Lowess	58.0%	78.9	Yes
MLM	58.0%	78.8	No
Quantile	57.8%	79.3	No

Given the limitation of a single independent variable, a lowess model fits the data well. We recall that the cubic piecewise models include an additional factor - the English subject differential - and other factors can be included as desired or necessary. With these points in mind, we consider that a lowess approach – whilst feasible – does not hold any benefits that the piecewise model approach does not convey.

A significant cause of unexplained residual variation is due to the numbers of entries made by pupils. Indeed, when number of entries is included in the simple model, the proportion of variance explained increases from 58% to 84%.