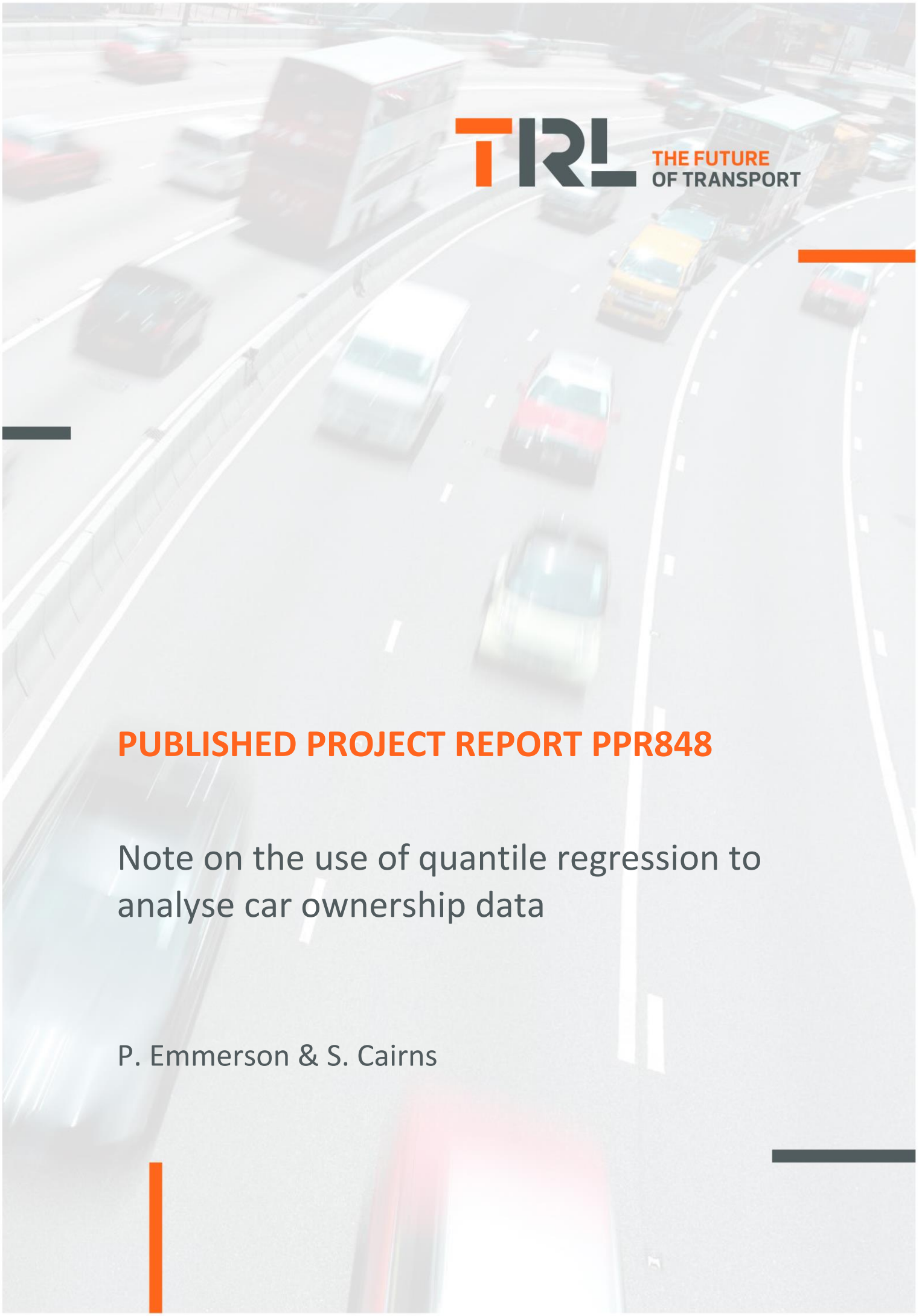

THE FUTURE OF TRANSPORT

PUBLISHED PROJECT REPORT PPR848

Note on the use of quantile regression to analyse car ownership data

P. Emmerson & S. Cairns

Report details

Report prepared for:	MOT research project (UKRC Energy Programme/EPSRC)	
Project/customer reference:	11111791	
Copyright:	© TRL Limited	
Report date:	November 2017	
Report status/version:	Final	
Quality approval:		
Katie Millard Project Manager		Caroline Wallbank Technical Reviewer

Disclaimer

This report has been produced by TRL Limited (TRL) as part of MOT research project (UKRC Energy Programme/EPSRC). Any views expressed in this report are not necessarily those of MOT research project (UKRC Energy Programme/EPSRC).

The information contained herein is the property of TRL Limited and does not necessarily reflect the views or policies of the customer for whom this report was prepared. Whilst every effort has been made to ensure that the matter presented in this report is relevant, accurate and up-to-date, TRL Limited cannot accept any liability for any error or omission, or reliance on part or all of the content in another context.

When purchased in hard copy, this publication is printed on paper that is FSC (Forest Stewardship Council) and TCF (Totally Chlorine Free) registered.

Acknowledgements

This analysis has been undertaken as part of the MOT project (EP/K000438/1), funded by the UK Engineering and Physical Sciences Research Council under the Research Councils UK Energy Programme, leading on from Research Councils UK Energy Programme research grant EP/J004758/1. This project has been led by Prof Jillian Anable at the University of Leeds, also involving TRL, University of the West of England, University of Bristol, University College London and University of Aberdeen. The work has been formally supported by the UK Department of Transport (DfT) and the former UK Department of Energy and Climate Change (DECC). This project includes use of Census data from the Office for National Statistics Crown copyright and database right 2012; and income data from the Office for National Statistics Crown copyright 2015. Project outputs can be found at http://www.MOTproject.net Analysis in this report is based on the version 7 project dataset. Grateful thanks to DfT, DVLA, DVSA, DECC, the project advisory board, the other project partners and other researchers who have worked on the project at TRL.

Table of Contents

Abstract

Quantile regression provides a way of investigating how the influence of individual variables changes across a distribution of observed output values. Specifically, in this report, it has been used to explore how the influence of income, population density, and the proportion of those aged over 65 impact on the distribution of car ownership values around an estimated mean, in areas where car ownership is relatively high or low. The approach adopted has involved use of a measure of car ownership generated through the MOT project, together with 2011 Census data and ONS data, at a Medium Super Output Area (MSOA) level.

This work has involved exploration of one output variable and three independent variables. However, the approach could be used for more complex data and models, and might be particularly useful for investigating the role of key determinants of car use, where the relative influence of the variables may vary more than for car ownership.

1 Introduction

This note describes exploratory work undertaken as part of the UKRC Energy Programme/EPSRC-funded MOT project. The project explored a range of different analytical techniques for understanding car ownership and use (see, for example, Yeboah et al 2016; Cairns et al 2017, www.MOTproject.net).

Following advisory panel input, and other work by the team using ordinary least squares regression and geographically-weighted regression techniques, the decision was made to investigate the potential of quantile regression for looking at levels of car ownership, drawing on use of the technique in other fields. Good summaries of the uses of quantile regression are available, see, for example, Cade and Noon (2003) for ecology and Koenker and Hallock (2001a) for econometrics. Other academic presentations are also available, for example, Baum (2013). This report describes a relatively simple application of the technique in order to explore its potential.

2 What is quantile regression?

Ordinary least-squares (OLS) regression estimates the impact of independent variables on the mean value of a dependent variable. The estimate of the variability around that mean value is derived from assumptions about the underlying error distribution and the general fit of the data. Such a regression approach can be susceptible to the influence of outliers and assumptions about the distribution of errors around the mean.

Quantile regression is a regression approach which provides a best-fit relationship not to the mean value of the dependent value, but to a specific quantile of the distribution around that mean value. In other words, regression results are calculated for a specific 'quantile' or percentage of the distribution of the dependant variable. Quantile regression can be undertaken for any value of the distribution, such as the median (50%), the 5%, or the 90% quantile. This approach is seen to have a number of advantages over the traditional least-squares approach in that:

1. The overall pattern of results produced is less susceptible to the influence of outliers. This is because the regression is based on an ordered dataset of the dependent variable, not the absolute values (i.e. results for individual quantiles may be affected by extreme results, but the overall pattern for all quantiles should be relatively unaffected.)
2. It is essentially independent of any assumptions about the error distribution. In fact, the regression process can provide insights into the likely error distribution because the individual quantile regressions can map out the output distributions – the variation in the intercept estimate at each quantile point for the mean values of the independent variables is probably best for estimating this.
3. Because the overall pattern of results is dependent only on the ordering of data, the results are invariant to monotonic transformations of the dependent variable. So, for instance, converting a logged dependent variable back to its untransformed state is easier than for the least-squares regression – the mean of a logged variable *is not*

the same as the unlogged mean of a variable; in contrast, the median of a logged variable *is* the same as the median of an unlogged variable.

4. It allows the independent variables to have different impacts on the various parts of the distribution of dependent variable values. So, for example, some variables may have greater impact at one or other of the extremes of the distribution, whilst having similar effects at the mean.
5. Arising from the last point, the technique can provide insights into explaining the distribution of values, which can be useful when, for example, we are interested in understanding those data points at the extremes of the distribution. Specifically, in this work, this approach can help to address the question - what is influencing those areas with higher/lower than average car ownership? See Cade and Noon (2003) for a discussion of the use of quantile regression for models with potential limiting factors.

There are a number of practical limitations of quantile regression:

1. It is difficult to transform the results from one spatial scale to another because, whilst suitable weighted means can be added together, medians cannot.
2. Quantile regression models are estimated by minimising the absolute errors of the dependent variable in the case of the median (50%) quantile. This requires a different optimisation procedure to that used with least-squares regression of the mean. In practice, there are model options available in various statistical packages but the statistics involved cannot be compared with those used for least-squares regression models.
3. Based on internet searches carried out to inform this task, the range of regression model forms available in existing statistical packages appears to be more limited than those available for least-squares regression. Linear quantile regression is available in the '*quantreg*' package in R. (The example dataset below was explored using this statistics package and the GENSTAT package. The results from the latter are very similar, although the default method of estimating standard errors is slightly more conservative with R).
4. The use of spatial analysis techniques in conjunction with quantile regression seems limited. Whilst the residuals from the quantile regressions can be tested for spatial stationarity within R, no statistical package appears to allow quantile regression for the calculation of Geographical Weighted Regression (GWR) or spatial lag/error models.

3 Data set

Two datasets were used to create the main output variable used in this model (cars per person): 2011 data containing information on privately registered cars in England and Wales in each Middle Super Output Area (MSOA)[1], from the MOT dataset, and Census data on the usual resident population in each area. Only one MSOA, in Swansea, was deleted because of an abnormally high number of cars per person, resulting in a dataset of 7,200 cases.

As part of other exploratory work undertaken for the MOT project, involving OLS regression modelling of car ownership, three significant variables were identified for use as independent variables in this exercise. These were:

- The natural logarithm of weekly net household income in £000 (generated from mean income estimates from ONS[2]) - LINC
- The square root of population density (generated from 2011 Census data, as persons per km) - SDEN
- The proportion of the population aged over 65 (generated from 2011 Census data) – POVER65

Figure 1 shows the distribution of observed values for the dependent variable 'cars per person' and Table 1 shows the summary statistics for the variables used in this study. The data suggests that all the data is significantly different from a 'normal' distribution and whilst the 'cars per person' shows a slight negative skewness - that is, it has a tail towards zero, the other variables have positive values – i.e. a slight tail towards the maximum values.

[1] The MSOA data set is based on an aggregation of the version 7 dataset for the LSOAs formed as part of the UKRC Energy Programme/EPSRC-funded MOT project. Each MSOA contains about 7,500 persons.

[2] http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-416744 (SmallAreaIncomeEstimatesdata_tcm77-420299.xls)

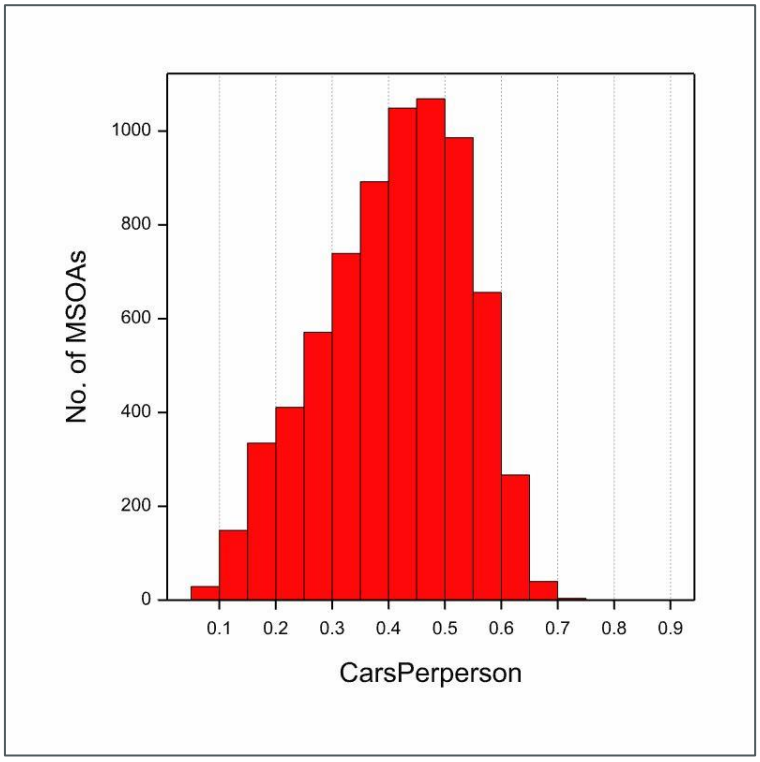

Figure 1: Distribution of observed values for 'Cars per person'

Table 1: Summary description of variables used

Variable	Cars per Person	LINC	SDEN	POVER65
Minimum	0.05099	5.473	2.386	0.005746
1st Quartile	0.32130	6.129	25.509	0.122790
Median	0.42312	6.227	49.536	0.161251
Mean	0.40980	6.282	49.007	0.161415
3rd Quartile	0.50750	6.426	66.590	0.197052
Maximum	0.87384	7.042	157.942	0.441747
Skewness	-0.336	0.192	0.532	0.333
Kurtosis (Normal = 0.0)	-0.543	-0.207	0.108	0.560

The quantile regression model was run for a number of different quantiles. This took the form of:

Cars per person (at a specified quantile) = Intercept + a*LINC + b*SDEN + c*POVER65

where 'Intercept', 'a', 'b', 'c' were parameters estimated by the statistical package which varied with the quantile being modelled.

4 Results

4.1 Hypothetical results

Before considering the results from the test data, it is useful to consider what results one might expect from quantile regression. If one assumes that the impact of any independent variable on the dependent variable at a given quantile is constant and equal to that for the mean value from a traditional Ordinary Least-Squares (OLS) regression, the estimated intercept values should mirror the output distribution of the dependent variable if the independent variables are input in a standardised form (i.e. with a mean of 0.0 and a standard deviation of 1.0).

To put this in context, Figure 2 shows the value of the standardised deviates[3] for a given quantile for the purely (hypothetical) forms of two common distributions (normal and log-normal). The values estimated for the intercept for each quantile would vary depending on the underlying distribution of the dependent variable, in this case 'cars per person'. For the majority of values in a normal distribution, there is an almost linear increase in deviate as the quantile increases but this does not hold for the extremes of the distribution. For the lognormal distribution, the approximately linear effect starts at very low quantile values but departs more severely from a linear relationship at high quantiles (0.9 and above). In the absence of any dependent variable impacts, a plot of the intercept values for each quantile that follows one of these distributions would be indicative of an underlying error distribution from that distribution. In other words, the increases in deviate values at either end of the distributions show that as you get further from the mean values, the impact of the 'tails' is greater for values of the intercept.

[3] Standardised Deviates represent the deviations from the mean values after taking account of the mean and standard deviation of the distribution of the variable, so that about 95% of the values would lie within ±2 standard deviates of the mean if the distribution was a normal distribution.

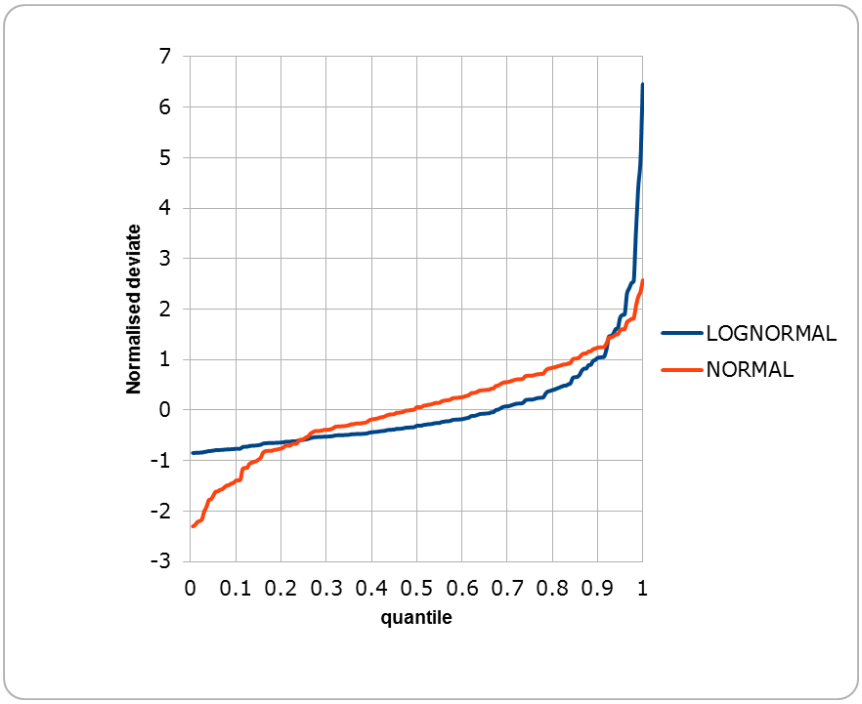

Figure 2: Variation in standardised deviate by quantile for two distributional forms

4.2 Standard regression results

Figure 3 shows the results estimated for the OLS solution using GENSTAT. The three independent variables (SDEN, LINC and POVER65) are all significant and together account for 82.7% of the variance in car ownership per person. No interactions between variables were modelled.

Regression analysis

Response variate: CarsPerPerson

Fitted terms: Constant + SDEN + LINC + POVER65

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	3	95.27	31.755666	11473.65	<.001
Residual	7195	19.91	0.002768		
Total	7198	115.18	0.016002		
Change	-1	-11.00	10.996309	3973.08	<.001

Percentage variance accounted for 82.7

Standard error of observations is estimated to be 0.0526.

Estimates of parameters

Parameter	estimate	s.e.	t(7195)	t pr.
Constant	-1.0720	0.0190	-56.33	<.001
SDEN	-0.0022385	0.0000276	-81.25	<.001
LINC	0.23033	0.00295	78.19	<.001
POVER65	0.8954	0.0142	63.03	<.001

Figure 3: Output from the GENSTAT OLS model

The relationships identified through this analysis predict that mean car ownership per person will increase as the natural logarithm of weekly income increases and as the proportion of an MSOA's population that is over 65 increases (the coefficient estimates are positive), but will decline as the square root of the density increases (the coefficient is negative). For example, a 10 percent point increase in the proportion of people in the MSOA older than 65 would increase the number of cars per person in the MSOA by around 9 percentage points (all other independent variables staying the same).

Figure 4 shows the distribution of the standardised residuals from this model.

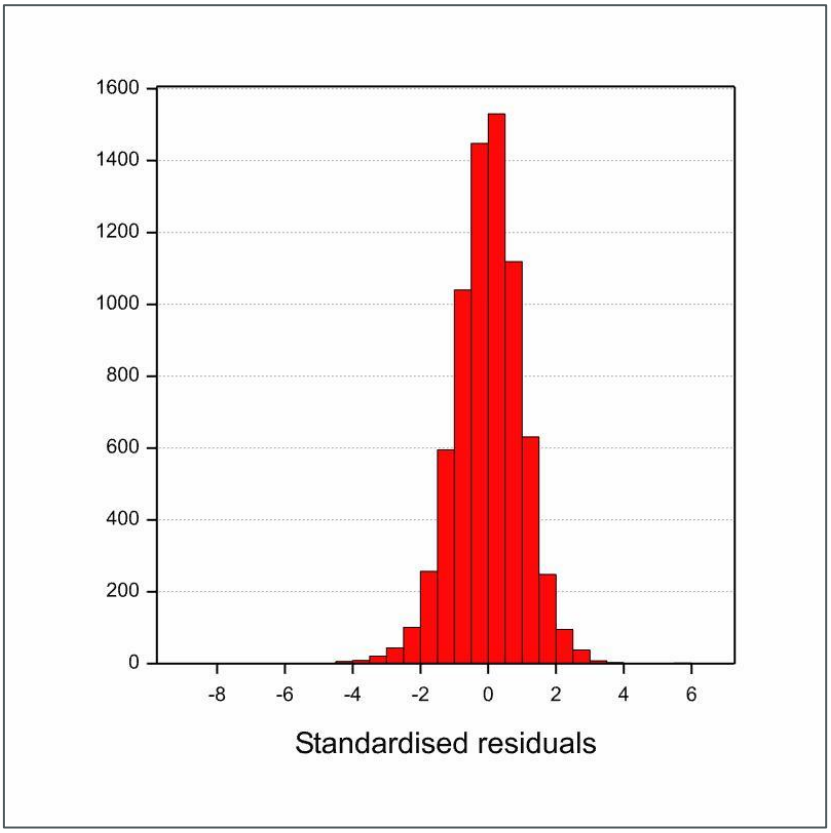

Figure 4: Distribution of standardised residuals from the OLS model

The chart shows that the residuals are approximately normal, but slightly skewed to the left. The statistical analysis output suggests that the error variance is greater for intermediate values of the dependent variable than for high or low values. (By way of comparison, a normal error distribution has constant error variance, whilst a Poisson or Gamma-like error distribution leads to the error variance increasing with the fitted value. A binomial distribution would have the greatest error variance in the intermediate values but this is usually only used for count data.) Hence, traditional assumptions about assumed error distributions do not hold for this model.

4.3 Quantile regression results

The same three variables were used to estimate the quantile estimates for the distribution of car ownership as used for the OLS regression above. Initially, the regressions were run using the raw independent variables (so that the intercept was negative). Following the work of Koenker and Hallock (2001b), further analysis was restricted to the relationship with the independent variables centred on their mean values so that the intercept represented the value of the dependent variable for that quantile at the mean values of the independent variables. It should be noted that this approach only affects the intercept value (which becomes positive) and the coefficients for the independent variables are the same as for the normal quantile regression. This approach enables a better interpretation of the intercept coefficients across the quantiles.

The summary option in the R-package '*quantreg*' library can automatically output the variation in the coefficients for each variable (and the intercept), as the quantile value

changes. The mean values of the coefficients for a given quantile are denoted by a black spot on the following graphs and the grey area gives an idea of the standard error of the estimate. (GENSTAT also produces estimates of the residual errors and standard errors but cannot display the OLS equivalent values on the same graph.)

4.3.1 General results

The three-variate model was run for quantiles of 5%, 25%, 50%, 75% and 95%. (The statistical packages actually estimates values of the coefficients for all quantiles). For each quantile, the procedure estimated a linear regression involving the three independent variates. Table 2 gives a summary of the results. The results show that the coefficient values for each parameter vary by quantile. Each of the parameters will now be discussed in turn.

Table 2: Summary of quantile regression results for representative quantiles

	Regression model coefficients (all independent variables mean centred)			
Quantile	**Intercept**	**SDEN**	**LINC**	**POVER65**
0.05	0.32427	-0.00223	0.20258	0.95621
0.25	0.37740	-0.00234	0.22025	0.92439
0.50	0.41152	-0.00226	0.24041	0.88448
0.75	0.44480	-0.00206	0.25935	0.83555
0.95	0.49099	-0.00177	0.29300	0.74106
Mean (OLS – mean-centred variables)	0.40980	-0.00224	0.23033	0.8954

Figure 5 shows the values of the intercept by quantile. Figure 5a shows the value of the intercept as the quantile changes. The actual values are plotted as black dots and the 90% confidence area is displayed as the grey area. The figure also shows the corresponding values for the OLS model as red lines for the mean and red dotted lines for the 90% confidence interval. The quantile distribution is normal-like (compare its shape to that for the normal distribution in Figure 2 but it is negative throughout. However, the quantile distribution is estimated where the value of all the independent values are zero (which is not meaningful in practise i.e. represents an area with a population density of zero, an average income of zero and 0% over 65).

A much more informative plot is displayed when a centered quantile regression is undertaken (c.f. Figure 5b) – that is, where the independent variables are input as deviations from their mean values. The intercept quantile distribution then represents the distribution of values at the mean of each independent variable and this is the form of model used in Table 2. From this chart, it is evident that the quantile distribution is approximately normal (and the median (50% quantile) and the mean estimate from an OLS regression are very similar) and has very tight confidence limits around it. That it should be

approximately normally distributed is not surprising given that the OLS residual errors have tended to be normally distributed.

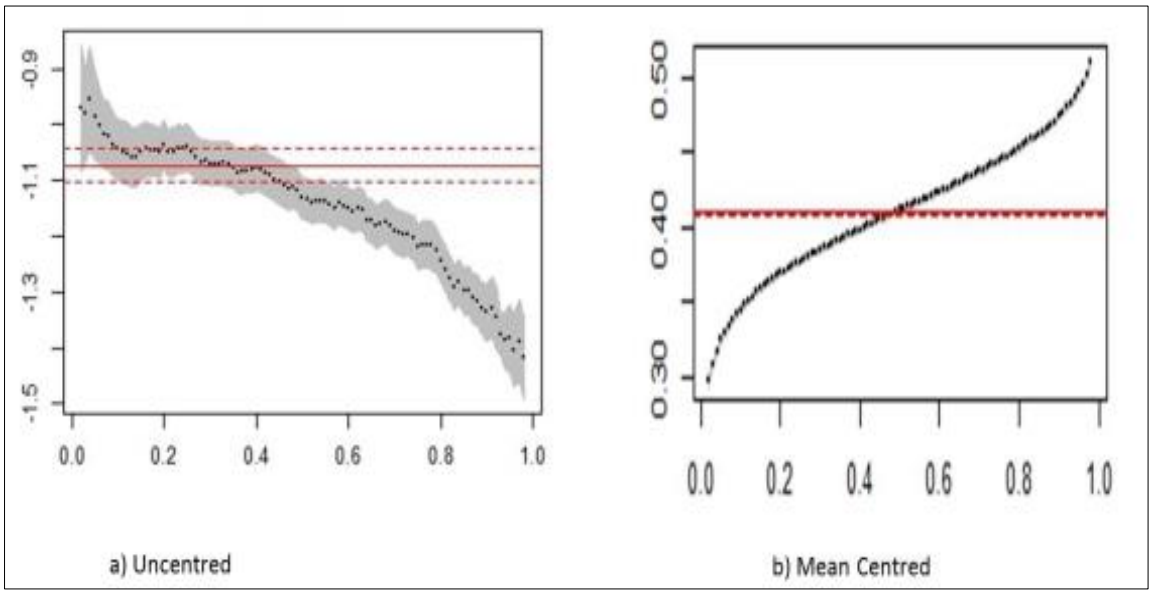

Figure 5: Quantile distribution of intercept values

(The x-axis gives the quantile as a proportion, whilst the y-axis gives the intercept value.)

4.3.2 Density

Figure 6 shows the equivalent quantile distribution for the density variable (in the form of the square root of the population density - SDEN). If the distribution of values of car ownership around the mean car ownership estimate were not affected by the actual value of the density parameter then the parameter values in the figure would be similar to that for the OLS regression whose value, is denoted by the red straight-line. The variation in the values for each quantile (shown as the black dots in the figure) do not follow a horizontal straight line which indicates that the independent variable is influencing the shape of the distribution of car ownership values, not just the mean value – it is influencing different parts of the distribution of car ownership values around the mean differently. While the impact of population density is to reduce mean car ownership by about -0.0023 cars per person for each increment of SDEN (and the impact on the median is similar to that on the mean), the impact of density is much less at the higher quantiles (-0.0018 at the 0.95 quantile). Thus the distribution of car ownership values around the mean is less sensitive to density than at high density levels, but has much less impact at lower levels of density (the slope is nearly flat).

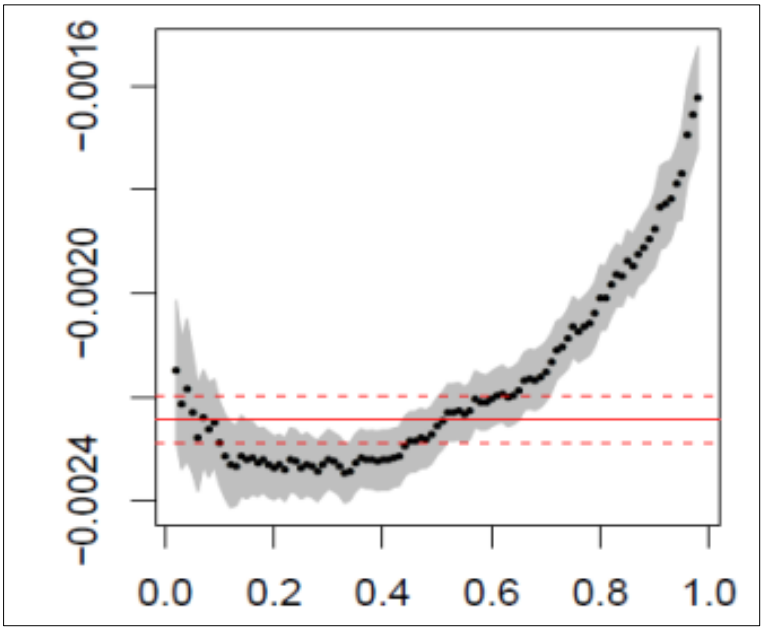

Figure 6: Quantile distribution of density coefficients (SDEN)

(The x-axis gives the quantile as a proportion, whilst the y-axis gives the coefficient value.)

Because the impact of density on car ownership varies with the quantile, it is likely that the *shape* of the distribution of predicted car ownership values around the mean will vary as density increases. The impact of density on the shape of the distribution is summarised in Figure 7 by plotting the absolute difference between two sets of percentile values – denoted as '45th percentile gap' on the y axis. If the shape of the distribution is unaffected by density then a plot of the difference between the 95th percentile and the median (Q95-Q50), and the difference between the median and the 5th percentile (Q5—Q5) should be constant and equal to approximately 0.11 (1.96*the residual standard deviation from the OLS regression). As the figure shows, these quantities do vary as density increases. In the case of the lower quantity (Q50-Q5), the change is quite small, but the upper quantity (Q95-Q50) increases markedly, indicating that the distribution becomes both more spread out and more skewed as densities increase.

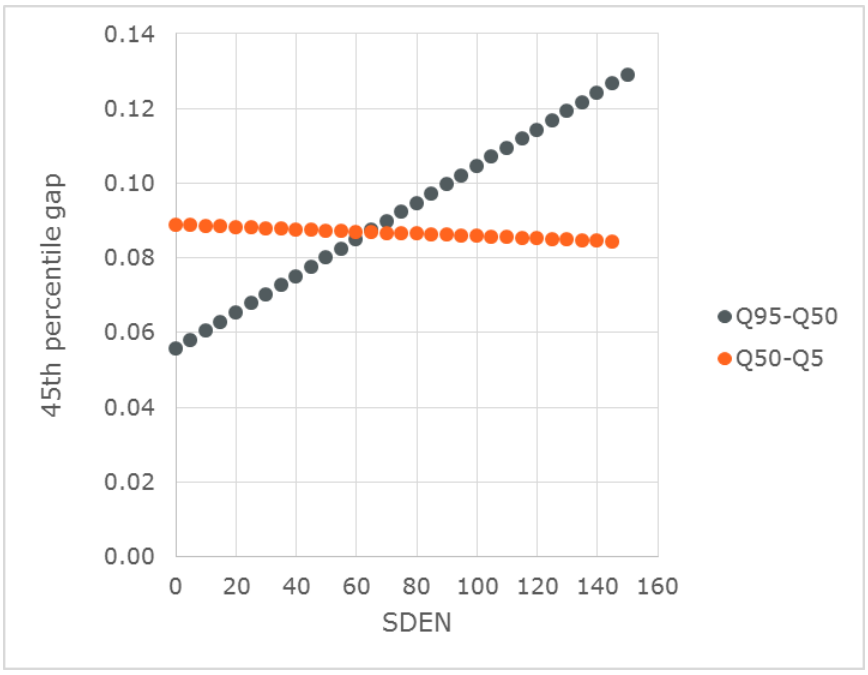

Figure 7: Impact of density (people/km) on the distribution of car ownership values

The impact that density has on the distribution of values of car ownership around the mean value can be seen in Figure 8. This graph shows the value of each quantile as population density increases, with all other variates at their mean values. In this figure the OLS estimate of the mean is very close to that of the median and so lies under the median (50%) line (i.e. the dotted line is not visible on the figure, as it is hidden by the green line). It should be noted that the x-axis is based on the untransformed value of density.

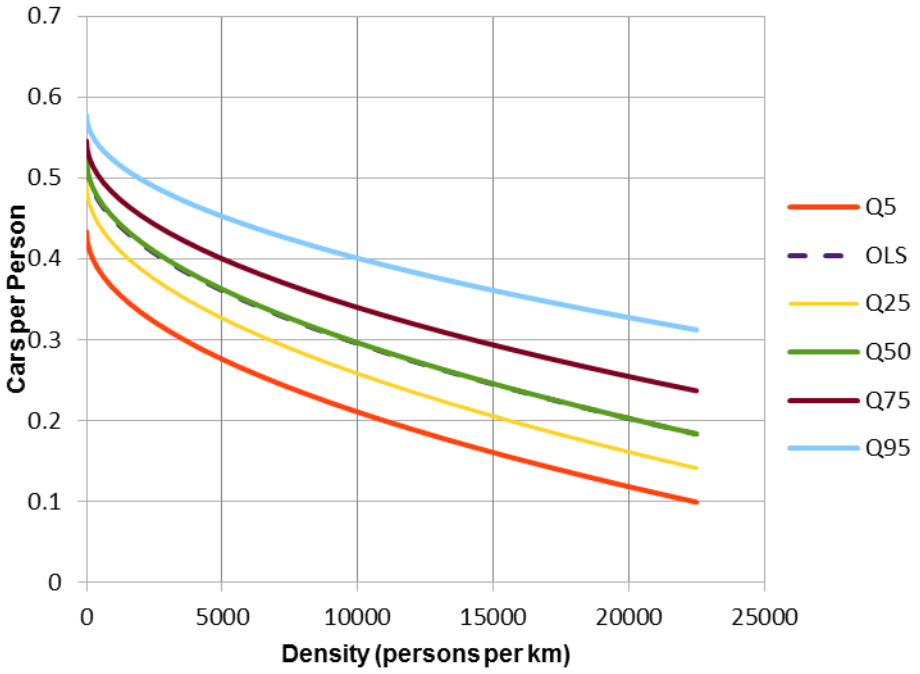

Figure 8: Variation in car ownership by quantiles by density

(The OLS value is hidden under the Q50 line, as explained in the text.)

4.3.3 Income

The modelled variation in quantile coefficients by income (actually logarithm of weekly income), shown in Figure 9, is quite different to that of the density variable. The variation in coefficient values with quantile is almost linear with the income coefficient across the whole range of quantiles, increasing in size as the quantile increases. This suggests that the distribution of car ownership values around the mean increases as income increases (increased heterogeneity), as well as the mean and median increasing with income (positive coefficients on the Y–axis). So the *distribution* of car ownership values around the mean becomes more sensitive to income as income increases.

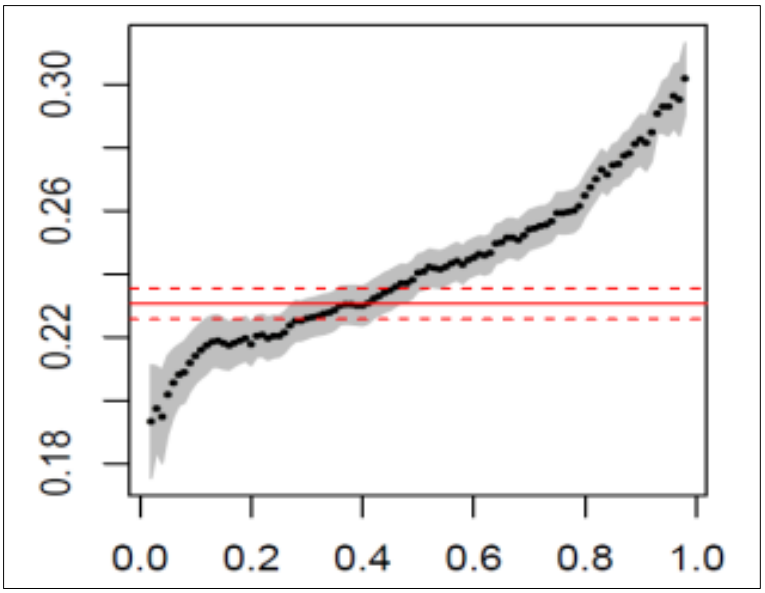

Figure 9: Quantile distribution of coefficients for income (LINC)

(The x-axis gives the quantile as a proportion, whilst the y-axis gives the coefficient value.)

Figure 10 summarises the impact of income on the *distribution* of car ownership values around the estimated means. Unlike the case with density, both halves of the distribution of car ownership values increase as income increases but the difference between the two 'halves' is much smaller, with the growth in the Q95-Q50 quantity only slightly greater than that for the Q50-Q5 – indicating that growth in the spread of the values is more noticeable than changes in the symmetry of the distribution, as income rises.

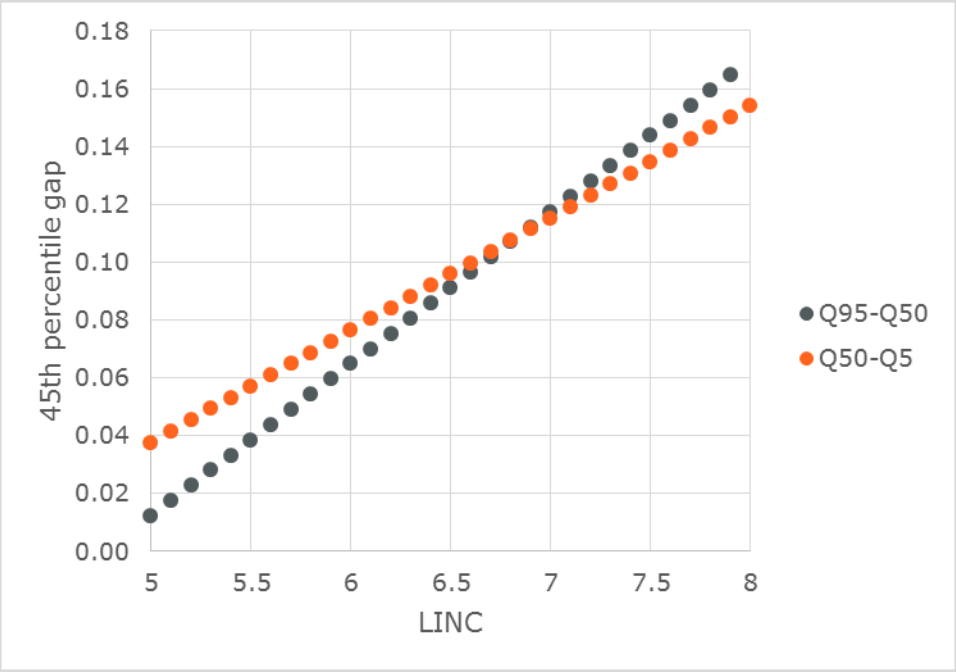

Figure 10: Impact of household income on the distribution of car ownership values

The impact of this variation on the distribution of car ownership values as log of income rises can be seen in Figure 11. One aspect of interest is the fact that the estimate of the mean car ownership level from the OLS regression (dotted line) does not lie near the estimate of the median car ownership level as estimated from the quantile regression (green line). Instead, OLS tends to give higher estimates of the mean than the quantile regression of the median.

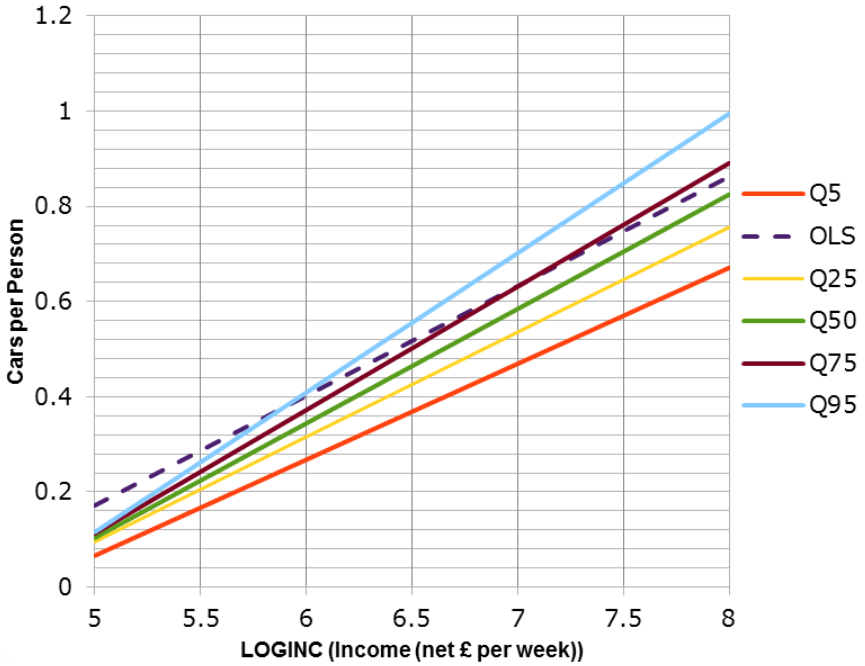

Figure 11: Variation in car ownership by quantiles by income

4.3.4 Proportion over 65

The trend in quantile coefficients for the third variable – the proportion of the population over 65, has a different form to the other two variables with the coefficients declining as the quantile increases (Figure 12). Thus, at high proportions of 65+ persons, the variability of car ownership values around the estimated mean is much less than at low proportions of 65+ persons.

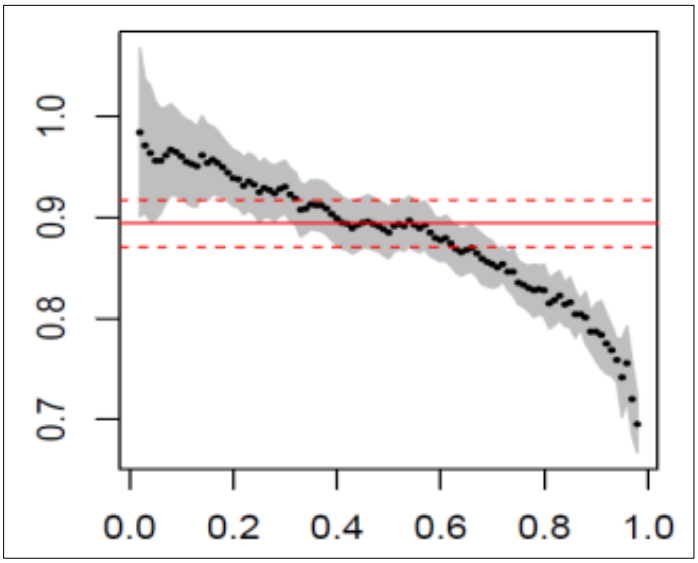

Figure 12: Quantile distribution of coefficients by the proportion of the population aged over 65 (POVER65)

(The x-axis gives the quantile as a proportion, whilst the y-axis gives the coefficient value.)

Figure 13 summarises the impact of the 'proportion over 65' on the distribution of car ownership values. In many ways, this is the mirror image of the income effect. As the proportion of over-65s increases, the spread of car ownership values becomes less pronounced and this is more so for the quantity 'P95-P50', leading to a bunching up of values in the upper half of the distribution. For example, the difference between the 95th percentile and the median when the proportion of over-65s is 0.40 is only 0.045 cars per person. When the proportion of over-65s is low, for example 0.1, this difference is 0.08 cars per person. Because both the '45th percentile differences' are reducing, this leads to a much sharper distribution of car ownership values around the estimated mean (high kurtosis) at high proportions of 65+ persons.

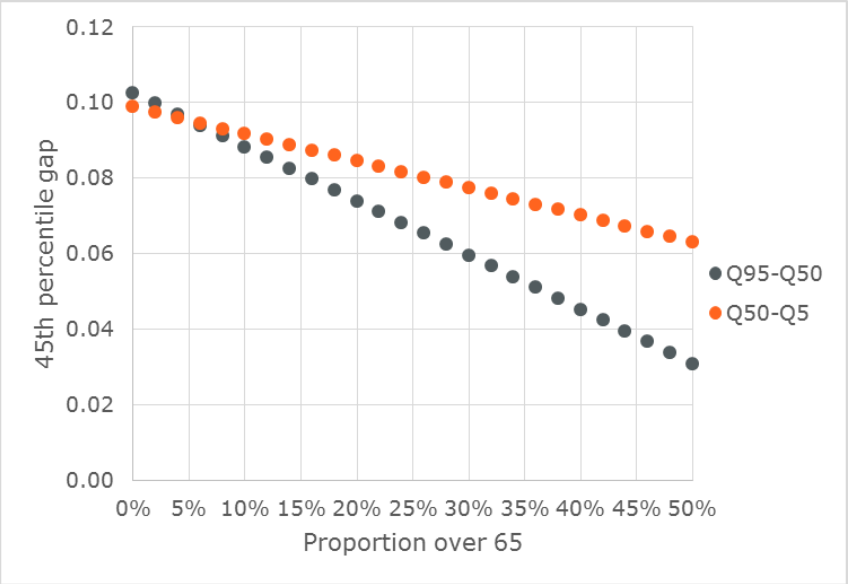

Figure 13: Impact of 'Proportion over 65' on the distribution of car ownership values

The impact of this variation on the distribution of car ownership estimates is shown in Figure 14 and is also quite different to that for the other two variables. The lines are all straight because the variable was entered into the regression models as an untransformed variable.

Unlike the other two variables, the variability of the estimates *decreases* as the proportion of the population over-65s increases (i.e. the lines get closer together); the mean estimate (dotted line) is always above the median value (green line) and for areas with a high proportions of over-65s, the mean is higher than the 95% quantile! This indicates that the OLS estimate is not a good predictor of the distribution car ownership values as the proportion of over-65s varies.

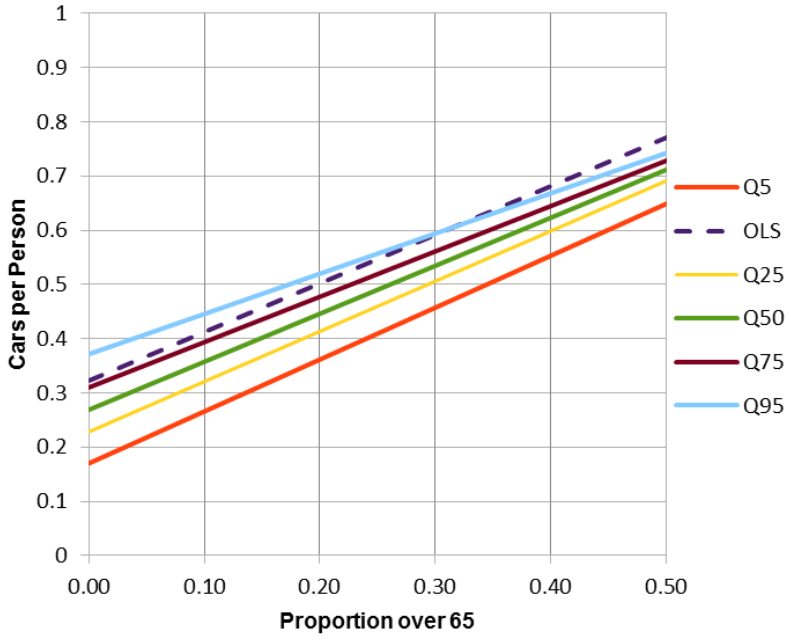

Figure 14: Variation in car ownership by quantiles by the proportion aged over 65

5 Discussion and summary

The implementation of quantile regression has highlighted a variety of additional characteristics in the relationship between car ownership per person and socio-economic variables at a MSOA level. This is despite the fact that the MSOA dataset did not show a great deal of non-normal distributional characteristics in the OLS regression of car ownership per person.

Whilst using the log of the income variable, the square-root of the density variable and the 'proportion of the population over 65' provide distributions where the mean and the median of the transformed variables are close to each other, each of the three variables investigated influences the ***distribution*** of car ownership estimates differently. This is highlighted best with the difference between the effect of income and the 'proportion of over-65s in the population' on the spread of the estimates (variance) of car ownership. Specifically, the work shows:

- The lower half of the distribution of car ownership values is not affected greatly by the density of the population, but as density increases the higher quantiles became increasingly sensitive to density, causing the estimated distribution of car ownership values for a given density value to have a tail to the right. As a result, small changes in population density are associated with relatively large changes to car ownership in the higher quantiles (e.g. the 90^{th} quantile).
- By way of contrast, as incomes rise, the distribution of car ownership values around the estimated mean value increases equally across the whole of the distribution. That is, the distribution of expected car ownership values looks approximately normal for all values of income but with the variability of the distribution increasing almost linearly as income rises.
- In the case of 'the proportion of the population over 65', as values for this variable increase, the variability of the car ownership distribution *decreases* (opposite to that of the income effect). Again the main impact is on the spread of the values rather than the symmetry of the distribution of the values.

Taken together it would seem that although the OLS regression can forecast the mean car ownership per person for a MSOA fairly accurately, and the residual error distributions are approximately normally distributed and show little heterogeneity relative to the fitted values, the quantile regressions show that the results based on simply estimating the mean and assuming a normal distribution of errors can hide a number of factors which affect the *distribution* of car ownership around these mean values. The quantile regression results show that each of the variables chosen for the OLS regression affects the *distribution* of the predicted car ownership values around the mean in different ways –these effects, in this dataset, show a reasonable conformity with OLS assumptions.

There are a number of possible reasons why the variation in the shape of the modelled distribution of car ownership with the three independent variables considered in this exercise could arise:

1. The form of the relationships in the OLS regression model may have been misspecified. The observation that the shape of the distribution widens with

increasing income, could be grounds for assuming that a change of error distributions to one where the variance increases as the mean values increase would be appropriate but, as the results have shown, the shape of the output distributions *decreases* with increasing 'proportion of over-65's so a simple change in the error distribution for an OLS regression is probably simplistic[4]. Note both variables affect the mean value in the same positive sense.

2. Another cause could be a misspecification of the regression model in terms what variables should be included. In other words, the complex interaction between the shape of the output distributions and the three independent variables in the current quantile models could be the result of interactions with variables not included in the model which have a significant impact on the distribution (and mean) of car ownership values. This is a plausible concern. Subsequently, work by the study team has shown that cross-sectional models for predicting car-ownership per person at a MSOA level can be more powerful if they include more than these three variables, and, when such models are developed, the three variables used here do not necessarily appear in the form used here or even at all (see for instance Emmerson et al, 2016, Table 1). It could be the influence of these missing variables that is causing the results shown here.
3. The final possibility is that these variables do have an individual impact on the shape of the distribution as well as on the mean car ownership level. One aspect mentioned in the ecological literature is that a variable may provide limiting growth only where other constraints are not at their limits. In these circumstances, one would expect a heavily skewed distribution with difference between the 95th percentile and the median to be much smaller than that between the 5th percentile and the median where a variable is influencing the upper limit – the long tail to the left represents cases where other variables are having greater impact (Cade & Noon, 2003, figures 2 and 3). In these cases, the quantile regressions of the highest quantiles (95th or 99th percentile for instance) will be the most useful in uncovering the processes involved. In our example, there are two potentially relevant cases. In the case of the proportion of 'over-65s', there is a greater reduction in the difference between the 95th percentile and the median than the reduction between the median and 5th percentile as the proportion of over 65s increases, indicating that there is a growing tail to the right. However, this effect is small compared with an increasing sharpness of the predicted car ownership distribution. This suggests that the proportion of elderly in the population may not be a limiting factor in the growth of high car ownership (because there is a wide spread of car ownership values at high proportions of 65+ persons). In contrast, in the case of population density (in the form of SDEN) the output distribution of car ownership values has an increasing tail

[4] The observed phenomenon that the error variance is greatest for intermediate values in the OLS regression may reflect the fact that although 'cars per person' is theoretically unbounded at its upper end (and bounded by 0.0 at the lower end), because cars are only owned by adults, and few adults own multiple cars, there is a theoretical constraint on higher values as well. As a result, the residual error distribution might be expected to show signs of the highest variance in the middle range of fitted values commensurate with assuming a binomial distribution..

to the right as densities increase, but little change in the lower half of the distribution across the range of densities, suggesting that, at higher densities, density is also not a limiting factor.

It should be remembered that the distribution of 'errors' around the mean values for the OLS regression of car ownership at a MSOA level was not that different from that of a normal distribution so there are implications for work which may have more extreme distributions; for example, dependent variables such as 'distance travelled per vehicle'. In particular, this approach could be used to identify and model units at the two extremes of the distribution (much higher and much lower than the expected mean) and so complement a study of the residuals from traditional regression models.

However, it should be borne in mind that currently, it seems that quantile regression cannot be combined with other forms of spatial analysis, especially Geographical Weighted Regression (GWR). This limits its applicability to modelling dependent variables such as car ownership, which we know, from previous work (Yeboah et al, 2016), to exhibit strong spatial correlations.

6 References

Baum C (2013) Quantile regression. EC 823 Applied Econometrics, Boston College, Spring 2013.

Cade B and Noon B (2003) A gentle introduction to quantile regression for ecologists. Frontiers in Ecology & Environment 2003: 1(8): 412 – 420, Ecology Society of America, 2003

Cairns S, Anable J, Chatterton T, Wilson E and Morton C (2017) MOToring Along: the lives of cars seen through licensing and test data. RAC Foundation, London.

http://www.racfoundation.org/assets/rac_foundation/content/downloadables/MOToring_along_Dr_Sally_Cairns_et_al_November2017.pdf

Emmerson, P, S Cairns , J Anable, S Ball, T Chatterton, J Barnes, E Wilson (2016) Using motor vehicle testing data to investigate spatial patterns of vehicle and energy use. Proceedings of the European Transport Conference. Barcelona, October 2016.
http://abstracts.aetransport.org/paper/index/id/5020/confid/21

Koenker Roger and Kevin F. Hallock (2001a) Quantile Regression. Journal of Economic Perspectives – Vol 15, no 1, Fall 2001, Pages 143-156. (Autumn 2001)

Koenker Roger and Kevin F. Hallock (2001b), Quantile regression: an introduction, http://www.econ.uiuc.edu/~roger/research/intro/rq3.pdf

Koenecker Roger (2015). Quantile regression in R: A Vignette. https://cran.r-project.org/web/packages/quantreg/vignettes/rq.pdf

Yeboah, G., Anable, A., Emmerson, P., Chatterton, T., Cairns, S., Ball, S. and Wilson, R.E.(2016) Exploring variance in car ownership levels at the local level: beyond income and density. Universities Transport Studies Group, Bristol, 6/1/16.

Note on the use of quantile regression to analyse car ownership data

Quantile regression provides a way of investigating how the influence of individual variables changes across a distribution of observed output values. Specifically, in this report, it has been used to explore how the influence of income, population density, and the proportion of those aged over 65 impact on the distribution of car ownership values around an estimated mean, in areas where car ownership is relatively high or low. The approach adopted has involved use of a measure of car ownership generated through the MOT project, together with 2011 Census data and ONS data, at a Medium Super Output Area (MSOA) level.

This work has involved exploration of one output variable and three independent variables. However, the approach could be used for more complex data and models, and might be particularly useful for investigating the role of key determinants of car use, where the relative influence of the variables may vary more than for car ownership.

Other titles from this subject area

MIS017	Understanding variation in car use: exploration of statistical metrics at differing spatial scales using data from every private car registered in Great Britain. Ball et al, 2016.
MIS018	Vehicle inspections – from safety device to climate change tool. Cairns et al, 2014.
PPR849	Impact of collinearity on the spatial analysis of car ownership and use. Emmerson et al, 2017.
PPR847	International experience of collecting and analysing technical inspection data for private cars. Millard et al, 2017.
PCN074	MOT data: what scope for understanding car ownership and use at a local level? Cairns et al, 2016.

TRL
Crowthorne House, Nine Mile Ride,
Wokingham, Berkshire, RG40 3GA,
United Kingdom
T: +44 (0) 1344 773131
F: +44 (0) 1344 770356
E: enquiries@trl.co.uk
W: www.trl.co.uk

ISBN 978-1-912433-09-4

PPR848