

## STAT0023 STATISTICS FOR PRACTICAL COMPUTING — ASSESSMENT 2 (2019/20 SESSION)

- Your solutions should be your own work and are to be submitted electronically to the course Moodle page by 12 noon on MONDAY, 27th APRIL 2020.
  - You can work either alone or in pairs for this assessment. It is up to you to form your own pairs. You MUST register your choices on Moodle by 12 noon on MONDAY, 30TH MARCH 2020, even if you choose to work alone.
  - If you choose to work in a pair, you will submit your joint work and you will be awarded the same mark.
  - Ensure that you electronically 'sign' the plagiarism declaration on the Moodle page when submitting your work.
  - Late submission will incur a penalty unless there are extenuating circumstances (e.g. medical) supported by appropriate documentation and notified within one week of the deadline above. Penalties, and the procedure in case of extenuating circumstances, are set out in the latest editions of the Statistical Science Department student handbooks which are available from the departmental web pages.
  - Failure to submit this in-course assessment will mean that your overall examination mark is recorded as "non-complete", i.e. you will not obtain a pass for the course.
  - Submitted work that exceeds the specified word count will be penalized. The penalties are described in the detailed instructions below.
  - Your solutions should be your own work. When uploading your scripts, you will be required to electronically sign a statement confirming this, and that you have read the Statistical Science department's guidelines on plagiarism and collusion (see below).
  - Any plagiarism or collusion can lead to serious penalties for all students involved, and may also mean that your overall examination mark is recorded as non-complete. Guidelines as to what constitutes plagiarism may be found in the departmental student handbooks: the relevant extract is provided on the 'In-course assessment 2' tab on the STAT0023 Moodle page. The Turn-It-In plagiarism detection system may be used to scan your submission for evidence of plagiarism and collusion.
  - You will receive feedback on your work via Moodle, and you will receive a provisional grade. *Grades are provisional until confirmed by the Statistics Examiners' Meeting in June 2020.*
- 

### Background and overview

Childhood obesity is a serious medical condition that affects children and adolescents. Children who are obese are above the normal weight for their age, height and gender. Obesity is prevalent worldwide; a [report by the World Health Organization \(WHO\)](#)<sup>1</sup> revealed that over 340 million children and adolescents aged 5-19 were overweight or obese in 2016.

Obesity is a major risk factor for a number of health consequences, including diabetes, cardiovascular disease, cancer, as well as psychological problems such as isolation and low self-esteem.

<sup>1</sup>Here and elsewhere, clicking on the blue text will take you to the relevant web site.

Obese children are likely to continue being obese during adulthood and are more likely to develop a variety of health problems as adults. A study tracking child obesity published in 2011<sup>2</sup> (among others) found that the probability of overweight children becoming overweight adults increases with a child's age.

In the UK, childhood obesity is monitored via the [National Child Measurement programme](#) (NCMP). Under this programme, measurements are taken on the height and weight of children in Reception class (aged 4 to 5) and year 6 (aged 10 to 11), to assess overweight and obesity levels in children within primary schools.

The NCMP programme runs every school year. Parents or carers of children eligible for measurement (i.e., attending state maintained schools at Reception Year and Year 6) receive a letter from local authorities informing them of the programme. They may choose to withdraw their child from the process. The height and weight of all eligible children (with consent) is then measured and submitted to NHS Digital and Public Health England (PHE) for further analysis.

NCMP uses the accepted method for diagnosing obesity in children by measuring their [Body Mass Index \(BMI\)](#), expressed as a percentile for their weight given a reference weight distribution (the WHO Growth Standards) of children of the same age, gender and height. A child is classified as overweight if their weight exceeds the 91st percentile and obese if they exceed the 98th.

The aim of the NCMP data collection is to study the factors associated with childhood obesity and understand appropriate mitigation strategies to tackle it. Interventions may be both direct and indirect. For example, a direct intervention may be to improve nutritional value of school meals; an indirect intervention may address underlying factors which are associated with obesity, such as socioeconomic indicators. A detailed understanding of the complex socioeconomic factors involved in child obesity is therefore paramount to any successful intervention. In this assignment, your aim is to build statistical models that will give you such an understanding.

The data provided for the analysis are a subset of the NCMP data, together with relevant socioeconomic indicators, spanning the period 2011 to 2017. The data have been obtained from the PHE database and contain annual information on numbers of Year 6 obese children in each of 326 Unitary Authorities (UAs — these are administrative districts) in England, along with corresponding socioeconomic indicators and the population size. PHE allows the use of these data free of charge in any format or medium, under the terms of the Open Government Licence v.3.0.

The data are provided as three separate files on the In-course assessment 2 tab of the STAT0023 Moodle page. The first, `obesity.csv`, contains a “cleaned” and anonymised version of the original obesity data: 2 232 observations (each row is an annual observation for a single UA) of Year 6 obesity counts and 14 additional covariates. Full details, including the anonymisation procedure (which includes small linear transformations of most variables) can be found in the Appendix to these instructions. The first 1 785 rows are complete, i.e., contain all values of obesity and covariates. The last 447 rows contain all values of the covariates, but “-1” for

---

<sup>2</sup>**Citation:** Starc G, Strel J. Tracking excess weight and obesity from childhood to young adulthood: a 12-year prospective cohort study in Slovenia. *Public Health Nutrition* 2011;14:49-55. doi:[10.1017/S1368980010000741](https://doi.org/10.1017/S1368980010000741). A copy of this paper can also be downloaded from the 'In-course assessment 2' tab of the STAT0023 Moodle page.

obesity counts. The second file, `UARegions.csv`, lists all the UA codes together with the wider region each UA belongs to (for example, UA EE01 is in the “East of England” region). The final file, `IndicatorNames.csv`, gives information about the variables in the PHE database.

Your task in this assessment is to carry out some data preprocessing to combine the datasets and then to use the data from the first 1785 records, to build a statistical model that will help you to:

- Understand the social, demographic and economic factors associated with variation in annual obesity Year 6 counts in each UA; and
- Estimate the Year 6 obesity counts for each of the 447 records where you don't have this information.

## Detailed instructions

You may use either R or SAS for this assessment.

1. Read the data into your chosen software package, define appropriate variable names (see the Appendix) and carry out any necessary recoding (e.g. to deal with the fact that ‘-1’ represents a missing value).
2. Combine the data from `obesity.csv` and `UARegions.csv` into a single dataset (a data frame if you're using R, a dataset in SAS), so that each row contains both the original data and the wider region that each observation corresponds to.
3. Carry out an exploratory analysis that will help you to start building a sensible statistical model to understand and predict annual obesity counts for each UA. This analysis should aim to identify an appropriate set of candidate variables to take into the subsequent modelling exercise, as well as to identify any important features of the data that may have some implications for the modelling. You will need to consider the context of the problem to guide your choice of exploratory analysis. See the ‘Hints’ below for some ideas.
4. Using your exploratory analysis as a starting point, develop a statistical model that enables you to *predict* annual obesity counts for each UA based on (a subset of) the other variables in the dataset, and also to *understand* the variation of obesity counts between different UA and across different years. To be convincing, you will need to consider a range of models and to use an appropriate suite of diagnostics to assess them. Ultimately however, you are required to recommend a single model that is suitable for interpretation, and to justify your recommendation. Your chosen model should be either a linear model, a generalized linear model or a generalized additive model.
5. Use your chosen model to predict the obesity counts for each UA and year where this information is missing, and also to estimate the standard deviation of your prediction errors.

Submission for this assessment is electronic, via the STAT0023 Moodle page. You are required to submit three files, as follows:

- A report on your analysis, not exceeding 2500 words of text plus two pages of graphs and / or tables. The word count includes titles, footnotes, appendices, references etc. — in fact, it includes everything except the two pages of graphs / tables. Your report should be in three sections, as follows:
  - I Describe briefly what aspects of the problem context you considered at the outset, how you used these to start your exploratory analysis, and what were the important points to emerge from this exploratory analysis.
  - II Describe briefly (without too many technical details) what models you considered in step (3) above, and why you chose the model that you did.
  - III State your final model clearly, summarise what your model tells you about the factors associated with variation of obesity counts in each UA per year, and discuss any potential limitations of the model.

Your report should not include any computer code. It should include some graphs and / or tables, but only those that support your main points. **Graphs and tables must appear on separate pages.**

In addition to your data analysis, **all pairs must include an additional page at the end of their report where each pair member briefly describes their contribution to the project.** You will need to agree this in your pairs before submitting the report. If both pair members agree that they contributed equally then it is sufficient to write a single sentence to that effect, or alternatively you are very welcome to describe your own personal contribution to the project. Note that this page will not be marked, nor will different marks be allocated to different pair members based on this. The purpose is to encourage you all to be mindful about contributing to this piece of group-work.

Your report should be submitted as a **PDF file named as #####\_rpt.pdf, where ##### is your group ID, with any spaces replaced by underscores (IMPORTANT!!!).** For example, if your group ID is "ICA2 C", your report should be named ICA2\_C\_rpt.pdf.

- An R script or SAS program corresponding to your analysis and predictions. Your script / program should run *without user intervention* on any computer with R or SAS installed, providing the files obesity.csv, UARegions.csv are present in the current working directory / current folder. When run, it should produce any results that are mentioned in your report, together with the predictions and the associated standard deviations. **The script / program should be named #####.r or #####.sas as appropriate, where ##### is your group ID with underscores instead of spaces.**

You may not create any additional input files that can be referenced by your script; nor should you write any code that requires access to the internet in order to run it. If you use R however, you may use the following additional libraries if you wish (together with other libraries that are loaded automatically by these): mgcv, ggplot2, grDevices,

RColorbrewer, lattice and MASS. You may not use any other libraries except those that are loaded automatically when you start R with a standard configuration (if in doubt, test your script using the UCL Remote Desktop).

- A text file containing your predictions for the 447 observations with missing counts. **This file should be named #####\_pred.dat, where ##### is your group ID with underscores instead of spaces.** The file should contain three columns, separated by spaces and with *no header*. The first column should be the record identifier (corresponding to variable ID in file obesity.csv); the second should be the corresponding count prediction, and the third should be the standard deviation of your prediction error.
- **NOTE:** if you work in pairs, **both members of a pair must confirm their submission on Moodle before the submission deadline.**

## Marking criteria

There are 75 marks for this exercise. These are broken down as follows:

**Report: 40 marks.** The marks here are for: displaying awareness of the context for the problem and using this to inform the statistical analysis; good judgement in the choice of exploratory analysis and in the model-building process; a clear and well-justified argument; clear conclusions that are supported by the analysis; and appropriate choice and presentation of graphs and / or tables. The mark breakdown is as follows:

**Awareness of context: 5 marks.**

**Exploratory analysis: 10 marks.** These marks are for (a) tackling the problem in a sensible way that is justified by the context (b) carrying out analyses that are designed to inform the subsequent modelling.

**Model-building: 10 marks.** The marks are for (a) starting in a sensible place that is justified from the exploratory analysis (b) appropriate use of model output and diagnostics to identify potential areas for improvement (c) awareness of different modelling options and their advantages and disadvantages (d) consideration of the social, economic and demographic context during the model-building process.

**Quality of argument: 5 marks.** The marks are for assembling a coherent 'narrative', for example by drawing together the results of the exploratory analysis so as to provide a clear starting point for model development, presenting the model-building exercise in a structured and systematic way and, at each stage, linking the development to what has gone before.

**Clarity and validity of conclusions: 5 marks.** These marks are for stating clearly what you have learned about how and why obesity counts vary across years and UAs, and for ensuring that this is supported by your analysis and modelling.

**Graphs and / or tables: 5 marks.** Graphs and / or tables need to be relevant, clear and well presented (for example, with appropriate choices of symbols, line types, captions,

axis labels and so forth). There is a one-slide guide to 'Using graphics effectively' in the slides / handouts for Lecture 1 of the course. **Note** that you will only receive credit for any graphs in your report if your submitted script / program generates and automatically saves these graphs, appropriately labelled, when it is run.

**Note** that you will be penalised if your report exceeds EITHER the specified 2500-word limit or the number of pages of graphs and / or tables. Following [UCL guidelines](#), the maximum penalty is 7 marks, and no penalty will be imposed that takes the final mark below 30/75 if it was originally higher. Subject to these conditions, penalties are as follows:

- More than two pages of graphs and / or tables: zero marks for graphs and / or tables, in the marking scheme given above.
- Exceeding the word count by 10% or less: mark reduced by 4.
- Exceeding the word count by more than 10%: mark reduced by 7.

In the event of disagreement between reported word counts on different software systems, the count used will be that from the examiner's system. The examiners will use an R function called `PDFcount` to obtain the word count in your PDF report: this function is available from the Moodle page in file `PDFcount.r`.

**Coding: 15 marks.** There are 4 marks here for reading the data, preprocessing and setting up variable names correctly and efficiently; 7 marks for effective use of your chosen software in the exploratory analysis and modelling (e.g. programming efficiently and correctly); and 4 marks for clarity of your code — commenting, layout, choice of variable / object names and so forth.

**Prediction quality: 20 marks.** The remaining 20 marks are for the quality of your predictions. **Note**, however, that you will only receive credit for your predictions if your submitted `#####_pred.dat` file is identical to that produced by your script / program when it is run: if this is not the case, your predictions will earn zero marks.

For these marks, you are competing against each other. Your predictions will be assessed using the following score:

$$S = \sum_{i=1}^{447} \left[ \log \sigma_i + \frac{(Y_i - \hat{\mu}_i)^2}{2\sigma_i^2} \right].$$

where:

$Y_i$  is the actual Year 6 obesity count (which we know) for the  $i$ th prediction;

$\hat{\mu}_i = \hat{\mathbb{E}}(Y_i)$  is your corresponding prediction;

$\sigma_i$  is your quoted standard deviation for the prediction error.

The score  $S$  is an approximate version of a *proper scoring rule*, which is designed to reward predictions that are close to the actual observation and are also accompanied by an accurate assessment of uncertainty (this was discussed during the Week 10 lecture, along with the

rationale for using this score for the assessment). Low values are better. The scores of all of the students in the class (and the lecturer) will be compared: students with the lowest scores will receive all 20 marks, whereas those with the highest scores will receive fewer marks. The precise allocation of marks will depend on the distribution of scores in the class.

If you don't supply standard deviations for your prediction errors, the values of the  $\{\sigma_i\}$  will be taken as zero: this means that your score will be  $-\infty$  if you predict every value perfectly (this is the smallest possible score, so you'll get 20 marks in this case), and  $+\infty$  otherwise (this will earn you zero marks).

## STAT0023 Assessment 2 — Hints

1. There is not a single 'right' answer to this assignment. There is a huge range of options available to you, and many of them will be sensible.
2. You are being assessed not only on your computing skills, but also on your ability to carry out an informed statistical analysis: material from other statistics courses (in particular STAT0006, for students who have taken it) will be relevant here. To earn high marks, you need to take a structured and critical approach to the analysis and to demonstrate appropriate judgement in your choice of material to present.
3. At first sight, the task will appear challenging. However, there is a lot of information that can guide you: look at some of the web links earlier in these instructions, and at other commentaries on obesity, to gain some understanding of what kinds of relationships you might look for in the data.
4. When building your model, you have two main decisions to make. The first is: should it be a linear, generalized linear or generalized additive model? The second is: which covariates should you include? You might consider the following points:

**Linear, generalized linear or generalized additive?** This is best broken down into two further questions, as follows:

- *Conditional on the covariates, can the response variable be assumed to follow a normal distribution with constant variance?* In this assignment, the response variable cannot be negative and it is an integer. Therefore, it cannot have exactly a normal distribution. However, you may find that the residuals from a linear regression model are *approximately* normal — and you may judge that the approximation is adequate for your purposes. The 'constant variance' assumption may also be suspect: for positive-valued quantities, it is common for the variability to increase with the mean. If this is the case here, you need to decide whether it varies enough to matter: you need to think about whether the effect is big enough that you can improve your predictions (and hence your score!) by accounting for it e.g. using a GLM. You might consider using your exploratory analysis to gain some preliminary insights into this point.

- *Are the covariate effects best represented parametrically or nonparametrically?* Again, your exploratory analysis can be used to gain some preliminary insights into this. You may want to look at the material from week 6, for examples of situations where a nonparametric approach is needed.

**Which covariates?** The data file contains a lot of potential covariates, some of which are more important than others. You have many choices here, and you will need to take a structured approach to the problem in order to avoid running into difficulties. The following are some potentially useful ideas:

- *Look at other literature on childhood obesity and associated socioeconomic factors.* What factors are considered to be the most important characteristics associated with obesity? Can these be linked to covariates for which you have information? Obviously, if you do this then you will need to acknowledge your sources in your report.
- *Which geographical variables?* Each observation has its corresponding UA as well as the wider region it belongs to. Choosing whether to include these (or which one), depends on whether your data contains enough resolution to warrant using UA as a (factor) covariate; similarly for the regions.
- *How to incorporate time?* In the STAT0023 module we have not studied models for time series. However, you can think about including time as an additional covariate: make sure you clearly think about and explain what the implications of this are.

You should not start to build any models until you have formed a fairly clear strategy for how to proceed. Your decisions should be guided by your exploratory analysis, as well as your understanding of the context.

5. Don't forget to look for interactions! For example, it may be that the different regions have different environmental conditions (some may be mostly urban, others mostly rural) that are not represented by the covariates in the present dataset; these differences in environmental conditions may lead to regional differences in the relationships between covariates and response.
6. You probably won't find a perfect model in which all the assumptions are satisfied: models are just models. Moreover, you should not necessarily expect that your model will have much predictive power: maybe the covariates in the data set just don't provide very much useful information about obesity counts. You should focus on finding the best model that you can, therefore — and acknowledge any deficiencies in your discussion.
7. To obtain the standard deviations of your prediction errors, you need to do some calculations. Specifically:
  - (a) Suppose  $\hat{\mu}_i = \hat{\mathbb{E}}(Y_i)$  is your  $i$ th predicted obesity level and that  $Y_i$  is the corresponding actual value.
  - (b) Then your prediction error will be  $Y_i - \hat{\mu}_i$ .
  - (c)  $Y_i$  and  $\hat{\mu}_i$  are independent, because  $\hat{\mu}_i$  is computed using only information from the first 1785 records and  $Y_i$  relates to one of the 'new' records.

- (d) The *variance* of your prediction error is thus equal to  $\text{Var}(Y_i) + \text{Var}(\hat{\mu}_i)$ .
- (e) You can calculate the standard error of  $\hat{\mu}_i$  in both R and SAS, when making predictions for new observations — see Workshops 6 and 9. Squaring this standard error gives you  $\text{Var}(\hat{\mu}_i)$ .
- (f) You can estimate  $\text{Var}(Y_i)$  by plugging in the appropriate formula for your chosen distribution — for example, if you're using a linear model then this is just the error variance estimate, whereas if you're using a Poisson distribution (which is a possibility when the response variable is a count) then  $\hat{\text{Var}}(Y_i) = \hat{\mu}_i$ .
- (g) Hence you can estimate the standard deviation of your prediction error as  $\hat{\sigma}_i = \sqrt{\hat{\text{Var}}(Y_i) + \text{Var}(\hat{\mu}_i)}$ . In fact, for the case of linear models this is exactly the calculation that is used in the construction of prediction intervals (see your STAT0006 notes or equivalent).
8. Larger UAs will tend to have more obese children simply because they have bigger populations. You may therefore think that it is more sensible to model the effects of covariates upon the obesity *rates* (i.e. the proportions of obese children in the population) rather than the actual counts. However, the assignment instructions tell you to model and predict obesity counts. One option here would be to fit models to the rates and then to derive the corresponding expressions for the counts (since the count is equal to the rate times the population size). If you use a Poisson or quasiPoisson GLM or GAM however, there is a more elegant approach. To illustrate, suppose you use a GLM in which the obesity count  $Y_i$  has a Poisson distribution — with mean  $\mu_i$  say, where  $\mu_i$  potentially depends on the covariates. If the corresponding population size (i.e. value of popCount) is  $P_i$  then the obesity rate is  $Y_i/P_i$ , which has mean  $\mu_i/P_i = \lambda_i$  say and which does *not* have a Poisson distribution (notice, for example, that it can take non-integer values). Notice, however, that  $\mu_i = P_i\lambda_i$  so that

$$\log \mu_i = \log P_i + \log \lambda_i . \quad (1)$$

If you fit a GLM or a GAM to the counts  $\{Y_i\}$  therefore, then you could use a log link function and include a fixed term  $\log P_i$  in the linear predictor. In the GLM case, this gives a model of the form

$$\log \mu_i = \log P_i + \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} . \quad (2)$$

Comparing (2) with (1), you can now see that fitting the Poisson model (2) to the counts  $\{Y_i\}$  is equivalent to fitting the model

$$\log \lambda_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}$$

to the obesity rates  $\{Y_i/P_i\}$  — so this gives you a way to model the effects of covariates on these rates directly. The term  $\log P_i$  in (2) is called an *offset*. Both R and SAS allow the specification of offsets when fitting GLMs — look at the help system for details.

## Appendix: the obesity.csv dataset

### Data sources and pre-processing

The data provided in obesity.csv are derived from the [PHE fingertips database](#), which refers to the variables as 'indicators'. The data have been reduced to exclude covariates with missing entries, reducing both the total number of observations and well as the number of measurements available per observation. Some small linear transformations and anonymisation were also carried out, to prevent you from being able to identify the actual values that you're supposed to be predicting from information available online (by using linear transformations for this purpose, we ensure that your models will be exactly equivalent to models fitted on the original data).

The preprocessing steps are as follows:

1. From the [Fingertips database](#), all UA-level indicators<sup>3</sup> for the 'Physical activity' and 'Wider determinants of health' profiles were downloaded *except* a couple of employment indicators that contained conflicting information.
2. Some uninteresting variables (for example representing statistical summaries calculated by Public Health England) were removed from the data.
3. Obesity counts and population sizes were calculated from the rate information in the raw data (e.g. it is common to report health statistics in terms of 'rate per 1000 population', and this is how some of the PHE data were originally provided).
4. The UA names were anonymised by assigning codes e.g. 'EE01', 'EE02' randomly to the UAs within each region.
5. In cases where multiple records were available for a single UA in a given year, the records were averaged to give a single annual average value for the response variable and each covariate.
6. Data for years before 2011 and after 2017 were discarded, because not all variables have data before 2011 and after 2017.
7. In the resulting dataset, variables with more than 10% of missing values were discarded. After this, rows with *any* missing values were discarded.
8. A sample of roughly 80% of the records was selected for use in the 'model building' part of the assessment (this will be referred to as 'Group 1' below), with the remaining 20% used for 'prediction' ('Group 2'). This was done in such a way that the two samples were non-overlapping but had very similar distributions of all potential covariates. Specifically:
  - (a) For each region, 80% of the observations were sampled at random, without replacement, as candidates to use in Group 1; and the remaining 20% were allocated to Group 2.

---

<sup>3</sup>UA-level indicators correspond to an 'Area type' of 'District & UA (pre 4/19)' in the database — you may find this help if you want to find out more about what each variable represents.

- (b) For each of the numeric covariates in the data set, a Kolmogorov-Smirnov test was performed to test the null hypothesis that the underlying distributions in Groups 1 and 2 are the same.
- (c) The samples were accepted only if the  $p$ -values for *all* of the Kolmogorov-Smirnov tests were greater than 0.5. Otherwise, a new candidate sample was drawn in step (a) and the procedure was repeated.

The Kolmogorov-Smirnov test is used here as a convenient way to measure whether two distributions are roughly similar. Note, however, that the obesity counts were *not* included in this balancing exercise: this is because the performance of predictions would be artificially enhanced if they were included (for example, we would know that the mean obesity count for Group 2 is similar to that for Group 1). Note also that no attempt has been made to balance the groups in terms of *combinations* of the covariates.

- 9. The 'Group 2' records were placed at the end of the data table, with their obesity counts set to  $-1$ ; and a new ID variable was created so that each record has an ID number between 1 and 2 232.
- 10. Each of the numeric covariates was transformed linearly: the transformation was of the form  $ax + b$ , where  $a$  is a random number close to zero and  $b$  is a random number close to 1 (the values of  $a$  and  $b$  were different for each covariate). Also, the obesity and population counts were scaled by the *same* random number close to 1 (so as to preserve the proportion of obese individuals). This makes no difference to any models that you fit, because the corresponding regression coefficients will scale correspondingly; but it makes it more difficult for you to match the data with any information that you can find online.

## Description of variables

This section gives a brief description of each of the variables in `obesity.csv`. Note that, as provided, the covariate 'names' are actually numeric: they correspond to 'indicator numbers' in the PHE database. One of your first tasks is to set meaningful names for the covariates. More information about each of them can be found at the [PHE fingertips database](#): select either the 'Physical Activity' or 'Wider Determinants of Health' profiles, then click the 'START' button, go to the 'Definitions' tab, set 'County & UA (pre 4/19)' as the Area Type and choose your indicator from the drop-down menu.

Variable name	Description
ID	Record ID, from 1 to 2 232
Year	The year the measurement was taken
UA	Unique code for the UA
obesity	Number of measured Year 6 children which were classified as obese

*Continued on next page ...*

... continued from previous page

Variable name	Description
popCount	Total number of children measured in the given year and UA
10301	Pupil absence: percentage of half days missed by pupils due to overall absence (including authorised and unauthorised absence).
11202	Violent crime - violence offences per 1,000 population, based on police recorded crime data.
22401	Emergency hospital admissions due to falls in people aged 65 and over, directly age standardised rate per 100,000.
90356	The percentage of households in an area that experience fuel poverty: a household is considered to be fuel poor if they have required fuel costs that are above average (the national median level) and, were they to spend that amount, they would be left with a residual income below the official poverty line.
90360	Excess winter deaths index, measured as the ratio of extra deaths from all causes that occur in the winter months compared with the expected number of deaths, based on the average of the number of non-winter deaths.
90637	Violent crime - sexual offences per 1,000 population, based on police recorded crime data.
92309	% population aged under 18
92310	% population aged 65+
92899	Economic inactivity rate: the percentage of the population aged 16-64 years who are economically inactive (i.e. neither in employment nor unemployed according to the ILO definition [not employed, available to start work within two weeks, and actively sought employment within past four weeks]).
92924	Air pollution: fine particulate matter. Annual concentration of human-made fine particulate matter at an area level, adjusted to account for population exposure. Fine particulate matter is also known as PM <sub>2.5</sub> and has a metric of micrograms per cubic metre ( $\mu\text{g}/\text{m}^3$ ).
93111	Affordability of home ownership: ratio of median house price to median gross annual residence-based earnings
93350	Gender pay gap (by workplace location): the absolute difference between median gross hourly earnings (excluding overtime) of men and women as a proportion of median gross hourly earnings (excluding overtime) of men, presented as a percentage.

Continued on next page ...

... continued from previous page

Variable name	Description
93351	Average weekly earnings: median gross (before tax, National Insurance and other deductions) weekly earnings in pounds (£) of full and part-time employees paid through the PAYE system, excluding over-time.