

How can I use statistical modelling to predict the
impact of smoking during pregnancy on asthma
levels in England and Wales in 2023?

Max Richards

March 2023

King's College London Mathematics School

Abstract

This project uses mathematics to make predictions about asthma rates in England and Wales and specifically, the impact that MSP (maternal smoking in pregnancy) has on these rates. I will discuss how the mathematics behind certain modelling methods works and the process that takes place in creating statistical models. It also discusses numerous models of my own, establishing my results regarding asthma rates, smoking throughout pregnancy rates and childbirth rates and how they all influence one another. The first section of my project will discuss mathematical modelling and will look at statistical methodology and its application to my project. The second section will cover my own models, discussing what I had to consider, the calculations carried out and the models used. The investigation concludes with my results, and comments on their significance, as well as an evaluation of the project as a whole.

Contents

1	Introduction	4
2	Literature Review	5
3	Using mathematics to make predictions	8
3.1	How can mathematics be used to make predictions?	8
3.2	The modelling cycle	8
3.3	Forecasting methods	10
4	Forecasting methodology	14
4.1	Introduction	14
4.2	Key terms	14
4.3	Planning - PPDAC cycle	15
4.4	Making the models	15
4.5	Results	27
5	Analysis	28
5.1	Testing Significance	28
6	Conclusion	30
7	Personal Evaluation & Next Steps	31
8	Citations	33

1 Introduction

It is no secret that smoking has severe negative effects on people's health, and since the late 20th century increasingly harmful information regarding the act has been uncovered, yet even still it is one of the most purchased goods in many first world countries across the globe. We know the effects are severe, but to what extent does it impact the rates of asthma, one of the most widespread conditions across many first world countries today, in the children exposed to these harmful substances throughout pregnancy?

Predictions involve making statements about the future. This process can have a vast variety of purposes and is used in all aspects of society today, ranging from weather forecasts to expiry dates on all foods. We often make predictions to have a sense of control, if we can predict what will happen, we have a significantly greater chance of being able to control it [1], and hence, being able to alter the outcomes to have the greatest utility.

Mathematics is perhaps humanities' greatest tool when it comes to making predictions. Through mathematical methodology, we can produce more accurate, consistent, and reliable results and assess these results to a high standard. Using mathematics can help consider uncertainty and randomness and provide a sense of validity to work that may not be able to be attained through intuitive predictions. Developing mathematics has been the key driving force behind all predictions made every day across the world and has been the primary reasoning for the significantly improved quantity and quality of forecasts made in the past few decades.

This Investigation will involve the use of mathematical modelling methods to make predictions about the effect that smoking during pregnancy can have on asthma rates in England and Wales. It will create numerous statistical models which will allow us to make predictions on asthma levels in England and Wales in 2023, and how the rates of MSP influence this trend. Through continued improvement and adjustment, I will be able to develop and advance my predictions and will be using mathematics as the key tool for not only making the models themselves, but also to assure significance in my results and provide validity to my findings.

2 Literature Review

To carry out this project, a large amount of initial research needed to be carried out before I could begin making any form of models and predictions. This would involve research into the mathematics of modelling methods, research into prior case studies, and what statistics I should predict for. Each of these will be covered in my review of five key pieces of literature that I used in my initial research stage.

I started by reading an article on the WIRED website by New Yorker, Atlantic and New York times writer Vauhini Vara, titled ‘We Will literally predict their life outcomes’ [2]. The article dove into examples of projects and schemes that have attempted to make predictions about children’s lives, and the impact that these predictions can have both short and long term. It was extremely useful in this sense, it was able to provide examples of schemes that have tried to tackle comparable questions to mine, but also briefly evaluate effectiveness, mentioning the benefits that the projects had and/or the negatives. It expanded upon this, providing examples of new schemes that are currently in development, in which try to avoid these issues. This was extremely useful, as it enabled me to quickly see what others have failed to do in the past on similar projects. However, the article did not go into depth on what outcomes of life the individual schemes were trying to predict for, only highlighting very vague topics such as, ‘how happy they will be?’. No discussion was held on what was considered for this and the ways that they broke this colossal question down. Further to this the article did not focus on how they made predictions, and so lacked information in the methodology side. I had good reason to trust the integrity of this article given the authors affiliation with such large and prestigious companies, including The Atlantic, and the New York Times Magazine. Hence, reading this article allowed me to understand where projects of this theme needed to be headed to be of significance, and through evaluating the advantages and disadvantages of examples, I could directly learn from these, making sure to prevent the mistakes in my own project, whilst incorporating the good traits. The article also gave me a promising idea of what problems I may face and will need to overcome throughout my project.

I progressed in my research reading an article by NCT [3], a trusted and reliable source, which explored the impacts that smoking during pregnancy can have on the children in question. The article explored the various harms that MSP can inflict on the children, looking at the immediate and long term effects, and the separate ways the different exposures to ‘smoking’ can have an impact. It was especially useful to my research in this sense as it enabled me to gather an idea of the types of damage that smoking during pregnancy can cause. Furthermore, it enabled me to better comprehend the extent of the issue, which enabled me to grasp an idea of the significance my project can have on society. Further to this, there was also statistics in which provided the reader with even more of an idea of the extent of the issue being discussed. This was

meaningful to me as it assured me that my research was going to be significant as it is truly a huge issue in society today. However, the source was imperfect, it simply listed numerous issues that may arise yet, there was little information on how it influences asthma directly and had no mention of just how much of an impact MSP truly has. This source will be useful supplementary information throughout my project.

I then read an article titled ‘Maternal smoking during pregnancy and its influence on childhood asthma’, by Angela Zacharasiewicz of the University of Vienna [4]. The article was very empirically driven, and described the extent of the issue asthma proposes, as well as an in-depth breakdown of the causal relationship between MSP and asthma. Further to this it mentioned studies that had gone into this field of research, and provided an estimated increase in probability of children who were victims of MSP being born with asthma. All of this was extremely useful to my research, it provided lots of information on the matter, even mentioning prior studies. This was particularly impactful as it strongly encouraged me to make a model trying to look at this relationship and make predictions myself, as the project would clearly be of significance. This prompted further research into the question and led to me finding many asthma statistics on the UK government website which could be used to make a model myself. This was useful as it meant I was able to see what their results were, how they attained them and more importantly what I could do differently, that would allow me to better this model and improve the quality of forecasts in this field. However, the source was very one dimensional as it only really dove into the effects on each child individually and did not look at how it influenced asthma levels across nations. This was where my model was to be a slight improvement from prior studies, as I wanted to use data to identify the impacts of MSP on asthma rates on a much larger scale. I once again had good faith in the source’s credibility, coming from such an established and knowledgeable academic in whom has had many years of experience in the field.

Next, I read the article ‘Making predictions with regression analysis’ by Jim Frost, a regular columnist for the American Society of Quality’s statistics digest. [5] It explores the use of linear regressions in making predictions. The article not only explained what a linear regression is and how it works, but also broke it down step by step using an exemplar to demonstrate. Furthermore, it also introduced lots of key terms and ideas that I would have to come to grips with to make accurate and sufficient predictions and demonstrated a clear and organised way of making predictions. It also indicated very clearly the importance of finding a suitable amount of high-quality data and information, in order to create valuable predictions. However, the article did not properly discuss the mathematics behind the model and ignored parts of the modelling process, as it only really looked at how to deal with analysis. As my project was going to have to go through the entire modelling process, more information was needed, so there was a slight limitation to the source, however overall, it proved to be valuable and a great introduction to the world of making predictions via

statistical methods.

Following from the lack of information regarding the general mathematics and understanding of the modelling processes in the previous article discussed, I decided to read ‘the Art of statistics’ by Sir David Spiegelhalter [6] a book primarily focused on making predictions and the diverse ways in doing so. The book’s intention was to provide the reader with a foundation in the world of statistical forecasting, as it described methods, defined terminology, and exemplified how to perform these methods in a very categorical and almost chronological order. It was incredibly useful to myself, as it not only described and evaluated various modelling methods, but also taught me the ways in which I need to be thinking to make these predictions as well as how my results should be laid out in a clear, and coherent manner that is both meaningful to the reader and aesthetically pleasing. Furthermore, it also vigorously described how to evaluate the results from all these different methods, something incredibly useful to my research as this was not highlighted in my other sources. The book also held great credibility, being written by an incredibly-well established statistician of the University of Cambridge, with many years of experience working in statistical methodology to the highest level. However, despite it’s immense benefit, it did not give guidance on how and where to find data or mention any examples to do with my project topic. Nonetheless, this book is likely to be immensely helpful to me for the rest of my project and will be a piece of literature I keep returning to throughout.

In conclusion, these five sources will form the basis of my research. The literature allowed me to understand the scope of my project and the methodology necessary to carry out such predictions. ‘The art of statistics’ proved to be incredibly valuable for me in terms of making predictions and the ways in which this can be done, with great supplementary information coming from the article on regression analysis. Furthermore, the other three articles were especially useful in providing me with opportunities to view prior work that has taken place in the fields of my research, and how I could use these to make progressions and advancements to better previous projects. Additionally, these articles were able to provide me with a great initial understanding of the impact that MSP has on children, and the impact it has on asthma rates in those children, allowing me to develop this further. Hence, I can say that it massively helped establish the foundations for my knowledge regarding asthma levels and the effect of MSP and how important it is to be tackling such issues. However, despite the variation in the topic of the sources, a common theme amongst all was the importance of making predictions in the real world, which was pleasant to read, as a goal of mine was always to create a project of significance to society.

3 Using mathematics to make predictions

3.1 How can mathematics be used to make predictions?

Mathematical modelling can be used in all aspects of life, to make predictions about circumstances given a mixture of inputs. This can be done in a variety of ways, with many different methods being used in various cases. These models may be deterministic whereby it gives the exact same output for a particular set of inputs no matter how many times you recalculate it [7], or they may be stochastic models, whereby randomness is involved [8]. In this project I will be aiming to create a deterministic model, that is exact and will depend simply on the inputs, and nothing else, with no random variables involved.

Further to the numerous types of models, there is also a countless number of modelling methods in which can be used to make predictions. This project will incorporate the use of linear regressions and multilinear regressions as well as time series analyses. In this section, I will discuss the mathematics behind the methods, and evaluate the benefits of each method to my project, as well as their limitations.

3.2 The modelling cycle

In order to create a reliable model, it is necessary to implement a step-by-step plan, of how the process of making predictions should, and will, be carried out. The PPDAC model proposed by Mackay and Olford [9] was introduced to me upon reading ‘the Art of statistics’ by David Spiegelhalter [6] and is defined as a sequence of steps that describe the statistical method [9]. This model can be summarised below in figure 1 [6].

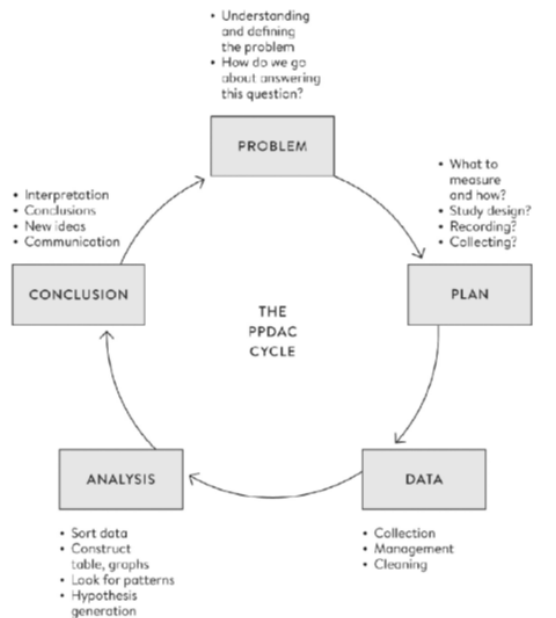


Figure 0.3
The PPDAC problem-solving cycle, going from Problem, Plan, Data, Analysis to Conclusion and communication, and starting again on another cycle.

Figure 1: The PPDAC problem-solving cycle

The First stage of any statistical inquiry is to determine what question you are going to be answering. Often this question will be highly specific and simplified to tackle part of a much larger question, that is extremely difficult to tackle itself. The next stage of the project should be to create a plan of how you will carry out the prediction process. This will then be followed by a high level gathering of data, from well-respected sources, concluding with the last stages of the prediction process itself. That being, according to this model, the analysis of our data and using it to make predictions, which we will evaluate, before sharing results in publications.

The main aim of any modelling process is to make predictions about something that can be related to, and understood, in the real world. However, this can often be problematic for many different models used today, in which are not necessarily going to always guarantee the exact outcomes, or make assumptions that are unattainable in real life, such as models involving negligible air resistance and friction. It is for these reasons that critics have risen to challenge mathematics' role in making predictions, as it is often deemed that they are a set of equations in which are unrealistic and massively oversimplified [10]. The reality is, mathematical models can be incredibly complex or remarkably

simple, and on a large variety of occasions the simple models produce outcomes which are extremely near if not identical to the outputs produced via the complex methods. Whilst there may be some truth in the idea that these more complicated models produce outcomes closer to the true value, they are often extremely difficult and time consuming to work with. Hence, when making predictions a key skill required is the ability to weigh up these factors when deciding what methods, to use to carry out the modelling process.

Furthermore, the aim of using mathematics in many models is to calculate relative probabilities for all the different possible outputs. The model should then be able to analyse these and produce the output with the highest probability of occurrence. This is what we will then take to be our output and will use this result to draw conclusions.

As mentioned previously, the purpose of any model is to make a valuable and accurate prediction about something. Whatever that something is, the main aim does not change. This has meant that over the years, mathematicians have continuously improved upon the models of their peers to improve estimates and better output predictions ever chasing the ‘perfect model’. The reasoning for this is due to there often being an element of uncertainty and error in the model. This is something that is unavoidable in most models and hence is why we are always able to see improvements being made.

The PPDAC model demonstrates clearly the methodical approach that needs to be taken to carry out predictions, and how making predictions through often imperfect mathematical approaches is a forever evolving process, which constantly improves through trial, and redevelopment. Despite this, there are often underlying assumptions being made that can oversimplify the project and in doing so, lead to sources of error and the models being unable to be applied to the real world. The key to carrying out the modelling process, is to determine relative probabilities of outcomes, considering sources of error, and by deciding on how the data analysis will be carried out considering factors that are ignored, and the complexity and manageability of different methods for the question you are trying to answer.

3.3 Forecasting methods

Following my initial research, the key method of statistical modelling I want to incorporate in my project was the use of linear regressions, a method in which has been incredibly significant in developing the field of statistical methodology and has proven to be an invaluable tool for data analysts across the world over the years, for numerous different purposes including business, academic study, sports analysis [11] and more.

A particular method of forecasting I wanted to use here was reference class forecasting, which as defined by conceptually is a method of predicting the fu-

ture by looking at similar past situations and their outcomes [12]. Using this strategy, I would be able to make predictions regarding my question which would be based off trends and relationships with time. The further back in time we go and hence the more data we gather, the more accurate the forecast is likely to be.

The key principle of linear regressions is looking at trends from data collected in the past, and identifying the best trend line, then make predictions which enables me to take vast amounts of data and use it to better manage reality – instead of relying on experience and intuition [11].

This section will discuss the mathematics behind linear regressions and how they work.

Linear regressions

The main idea behind linear regressions is that they make quantitative predictions about a variable (dependent/response variable) based on the input of other variables (independent/explanatory variable) [11]. Then by plotting the inputs, a line of best fit otherwise known as a regression line is plotted, often through using ordinary-least-squares (OLS), an idea proposed by Adrien-Marie Legendre and Carl Friedrich Gauss [6], which looks to make residuals of the regression line as small as possible, by summing the squares of all residuals of each point and plotting the regression line which results in the smallest sum. A residual is the distance between the regression line and our data point. When working with various explanatory variables, it would become far more difficult to plot the variables, however the OLS method still holds. The gradient of this mathematically calculated regression line is known as the regression coefficient, and this can be used in regression equations. This idea once again holds for one set of explanatory variables as well as multiple.

The general equation for all regression lines is given as follows; [13]

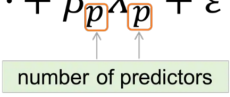
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$


Figure 2: The generalised formula for regression lines calculated via the OLS method

Y here represents our output of the function, or the dependent variable. Our values of β are constants in which are the matrices of linear coefficients for our independent variables, which shows how much a one-unit change in X changes Y. The magnitude of β expresses the effect changing X has on Y, and the sign of the coefficient gives you the direction of the effect. ϵ is the error term expressing how wrong our prediction is, and the X values represent our numerous independent variables [14].

To calculate the different values of β we must begin by calculating the residual using [14];

$$e = y - X\hat{\beta}$$

Figure 3: Formula for calculating the residuals of our regression

Where e is the residual, y is our true value and $X\beta$ is our fitted value, the value the model predicted.

We can then square the residual, and repeat this for all values, and then take the sum of all values [14];

$$\begin{bmatrix} e_1 & e_2 & \dots & \dots & e_n \end{bmatrix}_{1 \times n} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} e_1 \times e_1 + e_2 \times e_2 + \dots + e_n \times e_n \end{bmatrix}_{1 \times 1}$$

Figure 4: sum of squared residuals in matrix notation

This notation can be simplified to $e'e$.

We then want to differentiate our sum of squares term with respect to β' which can be done with the following steps [14];

$$\begin{aligned}
e'e &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\
&= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\
&= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}
\end{aligned}$$

$$\frac{\partial e'e}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta}$$

Figure 5: Derivative of sum of squared residuals

Now to find the value of β that minimises our sum of squares, we want to find the local minimum point, by setting derivative equal to zero and solving for β [14];

$$\frac{\partial e'e}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

Figure 6: Solving for best beta

This process should be repeated for each explanatory variable, to determine all the values of the β coefficients. Inputting these coefficient values into our regression equation alongside the explanatory variables enables us to calculate future values of Y.

4 Forecasting methodology

4.1 Introduction

We have seen in the previous section, the numerous intricacies we must account for and the considerations we must take to make forecasts. Further to this we also discussed the mathematical reasoning behind the forecasting methods which I will be using in this upcoming section.

The purpose of this EPQ project is to establish an understanding of forecasting methodology and to use the concepts and techniques learnt along the way to make predictions about the impact of smoking during pregnancy on asthma levels in the UK. This upcoming section will be covering that exact purpose precisely.

Furthermore, this section will explore uncertainty and sources of error, and account for these in models of my own.

4.2 Key terms

95% confidence intervals - The result of a procedure that, in 95 percent of cases in which its assumptions are correct, will contain the true parameter value. [6]

95% prediction intervals - A prediction interval is an estimated range of values that may contain the value of a single new observation, based on previous data. It is less certain than confidence intervals at the same percentage and includes a wider range of values. [15]

P-value - In linear regression, a P value indicates whether the relationship between an independent variable and the dependent variable is statistically significant while controlling for the other variables in the model. The closer the value is to zero, the higher the model's significance. [16]

f-statistic - These are a way of testing the significance of regression coefficients in linear regression models. The higher the f statistic the higher the model's significance. [17]

Correlation value (R) - A number between -1 and 1 that tells you the strength and direction of a relationship between variables. 1 indicates perfect positive correlation, -1 represents perfect negative correlation, 0 suggests no correlation at all. [18]

4.3 Planning - PPDAC cycle

Planning the numerous stages of my forecasting methodology using the PPDAC cycle;

Problem – The problem at large here, and what my models would be aiming to try and make predictions for, were the effects of MSP on asthma rates in the UK. A key goal with my project was to be able to deliver meaningful and worthwhile forecasts and attempting to tackle part of such a disastrous issue certainly provides that element of importance.

Plan – The goal was to use well-sourced data and attempt to make predictions using linear regressions on levels of asthma, pregnant smokers, child births and the relationships between all three.

Data – Majority of data was sourced from government office for national statistics [19] a trusted and reliable source of meaningful data. However numerous other sources were used for more data, all of which came from entrusted and verified sources.

Analysis – This will take place using linear regressions. Following analysis of data and obtaining results from these, a vast amount of time will be spent evaluating my forecasts, researching the sources of error in my methodology and the uncertainty in results.

Conclusion – A finalised write up of obtained results, including an evaluation of all methodology used and overall conclusiveness on the question topic.

4.4 Making the models

Regression 1:

My first stage of forecasting was an attempt to understand the relationship between time and number of live births each year. I decided to only consider live births and disregard still births as in 2020 it was found that on average in a random sample of 1000 births, 3.8 of these were still births [19], hence I disregarded the number of still births and took the statistics as regarding live births only and would assume that my findings from this data, could be generalised to all types of birth in England and Wales. For this a linear regression calculator was used from statskingdom.com , in which we used data from the government office for national statistics [20] to forecast the number of births in 2023.

The results obtained were as followed;

X - year	Y - No. live births	\hat{Y} (Predicted Y)	Residual
2012	729674	726452.7273	3221.2727
2013	698512	714637.8545	-16125.8545
2014	695233	702822.9818	-7589.9818
2015	697852	691008.1091	6843.8909
2016	696271	679193.2364	17077.7636
2017	679106	667378.3636	11727.6364
2018	657076	655563.4909	1512.5091
2019	640370	643748.6182	-3378.6182
2020	613936	631933.7455	-17997.7455
2021	624828	620118.8727	4709.1273

Figure 7: Data input into regression calculator and the estimated Y value for each year based off calculated regression line as well as residuals for each year.

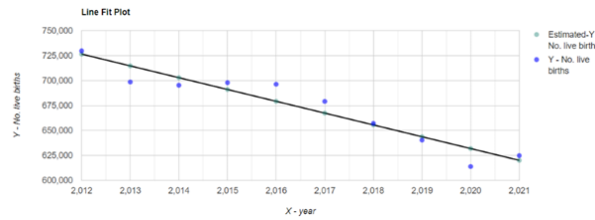


Figure 8: Graph showing the plotted data points and the regression line formulated.

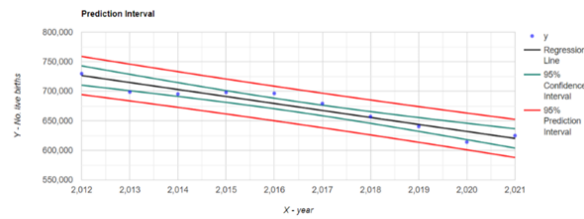


Figure 9: Graph showing the data points, regression line and the 95% confidence and prediction intervals.

The final regression line equation formulated was;

$$\hat{Y} = 24497976.65 - 11814.8727X$$

Figure 10: Calculated regression line equation.

Thus, substituting in a value of 2023 for X we obtain a prediction of 596,489 (to the nearest whole number) live births in England and Wales in 2023.

The f-statistic calculated was 79.1598 which indicates a remarkably high chance that the model is of significance. Additionally, the p-value obtained from this model was found to be 0.00002016 which again suggests an incredibly high chance that the regression model is of significance.

Furthermore, there was a correlation value, R, of -0.953 meaning there was an extraordinarily strong inverse relationship between X and Y.

Results:

Following the feedback from the statistical model, it was apparent to see that there is a clear strong negative correlation between time and number of live births each year, and the incredibly low p value and high f-statistic reassured me that the model was highly significant.

My result was an estimated 596,489 live births in the UK.

This clear trend could be due to a variety of reasons, for example the decline in fertility rates in women in the UK over time and the legalisation of abortions in 1967, [19] and hence it is not something that can be attributed to one specific factor.

Now to extend upon these findings I used national data regarding levels of asthma in the UK, from the centers for disease control and prevention website [21] which informed me that approximately 5.8% of all children in the UK have asthma. Using this statistic and multiplying it with our estimated number of live births found in our first model, we attain an estimated 34,596 (to the nearest whole number) of children to be born with asthma.

Regression 2:

The next model to be carried out would be to establish the relationship between the number of pregnant women who smoked, and time, and to establish a prediction for the proportion of women who will smoke during pregnancy in 2023. The data I used in this linear regression came from an article by statista.com. [22]

The results were as follows;

X - year	Y - % smokers	\hat{Y} (Predicted Y)	Residual
2012	12.9	12.5364	0.3636
2013	12.2	12.1727	0.02727
2014	11.7	11.8091	-0.1091
2015	11	11.4455	-0.4455
2016	10.7	11.0818	-0.3818
2017	10.8	10.7182	0.08182
2018	10.6	10.3545	0.2455
2019	10.4	9.9909	0.4091
2020	9.6	9.6273	-0.02727
2021	9.1	9.2636	-0.1636

Figure 11: Data input into regression calculator and the estimated Y value for each year based off calculated regression line as well as residuals for each year.

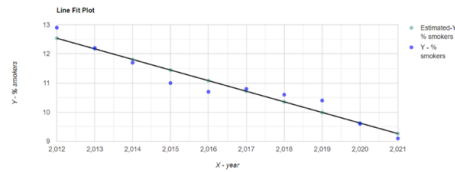


Figure 12: Graph showing the plotted data points and the regression line formulated.

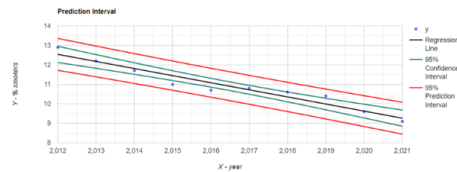


Figure 13: Graph showing the data points, regression line and the 95% confidence and prediction intervals.

The final regression line equation calculated was;

$$\hat{Y} = 744.1727 - 0.3636X$$

Figure 14: Calculated regression line equation.

Once again, taking this equation and substituting a value of 2023 for X we attain a prediction that 8.6% of all pregnant mothers' smoke through pregnancy.

The f-statistic calculated was 116.2228 meaning the model has an extremely high chance of being of significance. Additionally, the p-value obtained from this model was found to be 0.00000483 which also suggests an extremely high chance that the regression model is of significance.

Furthermore, there was a correlation value, R, of -0.9673 portraying an immensely strong inverse relationship between X and Y.

Results:

Following the feedback from the statistical model, it was apparent to see that there is a distinct strong negative correlation between time and percentage of women who smoke during pregnancy.

The significance of the model was assured via the remarkably low p-value and the incredibly high f-statistic.

My result was an estimated 8.6% of women who smoke(d) during pregnancy.

Regression 3:

A consideration I decided to take was the effects of multiple birth pregnancies. Using this information, I would be able to formulate a better prediction from the number of mothers who gave birth in my statistics, rather than just counting each birth as coming from a unique individual mother. For this I used data gathered from statista.com once again. [23]

The results were as follows;

X - Year	Y - Multiple birth %	\hat{Y} (Predicted Y)	Residual
2012	1.59	1.6151	-0.02511
2013	1.56	1.6016	-0.04161
2014	1.60	1.5881	0.01189
2015	1.61	1.5746	0.03539
2016	1.59	1.5611	0.02889
2017	1.58	1.5476	0.03239
2018	1.54	1.5341	0.005889
2019	1.54	1.5206	0.01939
2020	1.44	1.5071	-0.06711

Figure 15: Data input into regression calculator and the estimated Y value for each year based off calculated regression line as well as residuals for each year.

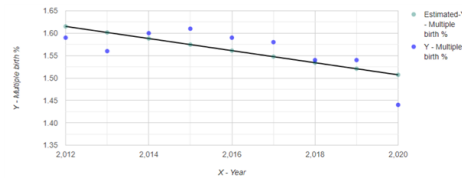


Figure 16: Graph showing the plotted data points and the regression line formulated.

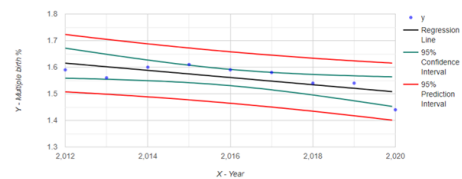


Figure 17: Graph showing the data points, regression line and the 95% confidence and prediction intervals.

The regression line equation calculated was;

$$\hat{Y} = 28.7771 - 0.0135X$$

Figure 18: Calculated regression line equation.

This leaves us with an estimated 1.4666% of all pregnancies in 2023 to be multiple carriages.

The regression was discovered to have a high chance of being statistically significant, given the f-statistic was found to be 7.2528, and the p-value 0.03095.

Further to this, the regression coefficient, R, was given a value of -0.7133 , which suggests a strong inverse relationship between X and Y.

Results:

Following the feedback from the statistical model, it was apparent to see that there is a distinct strong negative correlation between time and percentage of pregnancies in which multiple children are carried.

The significance of the model was assured via the low p-value and the high f-statistic.

My result for this particular regression was an estimated 1.4666% of pregnancies were carrying multiple children.

Calculations:

Now using the previous information and further research via an article from ReproductiveFacts.org [24] I wanted to estimate the expected number of children given that the mother was carrying multiple children. For this I carried out weighted averages using the following data;

$$P(\text{twins}) = 1/250$$

$$P(\text{triplets}) = 1/10,000$$

$$P(\text{quadruplets}) = 1/700,000$$

Given this information I deemed it suitable to take any probability of having quadruplets or more to be negligible as the effect on the average would be so minimal.

I then observed that you were 40 times more likely to have twins than triplets, hence doing the average of forty '2s' and one '3' you arrive at a final expected value of 2.024.

Now given the expected value of children, given the pregnant woman is carrying multiple children, the regression calculated number of live births in 2023, and the probability of any selected pregnant woman carrying multiple children, also calculated from a regression, I could establish the following simultaneous equations which would thus allow me to determine an estimate for the number of pregnancies in 2023, not just the number of live births.

$$0.014666(z) \times 2.024 = y$$

$$596489 - y = 0.985334(z)$$

Where z is the total number of pregnancies and y is the number of children born from multiple birth pregnancies.

Hence for 2023, we attain that (rounded to nearest whole numbers);

$$Z = 587,663$$

$$Y = 17,444$$

Repeating the simultaneous equations above for all years from 2012 – 2021 using the information from previous tables, I attained the following results;

Year	Expected number of pregnancies (z)	Expected number of children born in multiple carriage births (y)
2012	718,878	21,339
2013	688,177	20,428
2014	684,946	20,332
2015	687,527	20,409
2016	685,969	20,362
2017	669,058	19,860
2018	647,354	19,216
2019	630,895	18,727
2020	604,852	17,954
2021	615,583	18,273
2023	587,663	17,444

Figure 19: Screenshot of table showing values for z and y

Now using this information, we could invoke the results from our second regression, to attain a prediction for the number of women who smoked during pregnancy, looking distinctively at the cases where women carry multiple children and the cases where they only carried one child.

Multiple children:

$587,663 \times 0.014666 = 8619$ (expected number of pregnant women carrying multiple children)

$8619 \times 0.086 = 741$ (expected number of women who smoked during these multiple carrying births)

One child only:

$587,663 \times 0.985334 = 579,044$ (expected number of pregnant women carrying a single child)

$579,044 \times 0.086 = 49798$ (expected number of women who smoked during these single-carriage births)

Once again repeating these steps using data from table above and the smoking during pregnancy statistics from figure 18, resulted in the following updated table;

Year	Expected number of pregnancies (z)	Expected number of children born in multiple carriage births (y)	Expected number of women who smoked during multiple child carriage	Expected number of women who smoked during single child carriage	Total expected number of women who smoked during pregnancy	% Of women who smoked during pregnancy
2012	718,878	21,339	1360	91,375	92,735	12.9
2013	688,177	20,428	1231	82,726	83,957	12.2
2014	684,946	20,332	1175	78,963	80,138	11.7
2015	687,527	20,409	1109	74,519	75,628	11
2016	685,969	20,362	1076	72,322	73,398	10.7
2017	669,058	19,860	1060	71,199	72,259	10.8
2018	647,354	19,216	1006	67,613	68,619	10.6
2019	630,895	18,727	962	64,651	65,613	10.4
2020	604,852	17,954	852	57,214	58,066	9.6
2021	615,583	18,273	822	55,196	56,018	9.1
2023	587,663	17,444	741	49,798	50,539	8.6

Figure 20: Screenshot of a table showing results of numerous calculations.

According to cdc.gov, the national percentage of children aged 0-4 in whom have asthma in the UK is 2.0

Now according to a statistic from webmd.com, [25] children in whom were victims of their mother smoking throughout their pregnancy were 1.6 times more likely to have asthma than the children from mothers who did not smoke during pregnancy.

Hence multiplying the 2% chance that a child is born with asthma by 1.6, I concluded that when a mother smokes during pregnancy, the probability that her child gets asthma is 3.2%. So, I would multiply the number of children in that category by 0.032 to attain an estimate for the number of children in whom are born with asthma when their mothers smoke (in columns 5 & 6 in table below). However, when we are trying to attain estimates for the number of children with asthma in the cases where the mother does not smoke, we will only multiply our number of children born by 0.02 (in columns 7 & 8 in table below).

So, to improve on the initial prediction made above via a simple calculation I decided to break down the topic into categories and deal with them more specifically.

Year	Expected number of children born in multiple carriage pregnancies whereby the mother smoked	Expected number of children born in single carriage pregnancies whereby the mother smoked	Total expected number of children born subject to their mothers smoking during their pregnancies	Expected number of children born with asthma in multiple carriage pregnancies where the mother smoked	Expected number of children born with asthma in single carriage pregnancies where the mother smoked	Expected number of children born with asthma from multiple carriage pregnancies	Expected number of children born with asthma from single carriage pregnancies
2012	2753	91,375	94,128	88	2942	427	14,167
2013	2492	82,726	85,218	80	2647	409	13,562
2014	2378	78,963	81,341	76	2527	407	13,498
2015	2245	74,519	76,764	72	2385	408	13,549
2016	2178	72,322	74,500	70	2314	407	13,518
2017	2145	71,199	73,344	69	2278	397	13,185
2018	2036	67,613	69,649	65	2164	384	12,757
2019	1947	64,651	66,598	62	2069	375	12,433
2020	1724	57,214	58,938	55	1831	359	11,920
2021	1664	55,196	56,860	53	1766	365	12,131
2023	1500	49,798	51,298	48	1594	349	11,581

Figure 21: Screenshot of a table showing results of numerous calculations.

Now looking specifically at the proportions of all cases of children being born with asthma in which are caused/affected by smoking during pregnancy.

To calculate this, I did the following calculation.

$$\frac{\text{Sum of expected number of children born with asthma victims of MSP}}{\text{Sum of expected number of children born with asthma}}$$

Year	Expected number of children born with asthma whose mothers smoked during pregnancy/expected number of children with asthma each year
2012	0.2076
2013	0.1952
2014	0.1872
2015	0.1760
...	
2016	0.1712
2017	0.1728
2018	0.1696
2019	0.1663
2020	0.1536
2021	0.1456
2023	0.1376

Figure 22: Screenshot of a table showing results of numerous calculations.

The decimal value given in each row indicated the proportion of all asthma cases in that year that arose from mothers smoking during pregnancy.

Further to this I decided to investigate how the proportion of those who were born with asthma from mothers who smoked during pregnancy: the number of live births each year, had changed over time.

To calculate this, I did the following calculation.

$$\frac{\text{Sum of expected number of children born with asthma from mothers who smoked}}{\text{Expected number of live births}}$$

Year	Expected total number of people with asthma whose mothers smoked during pregnancy/number of live births
2012	0.004153
2013	0.003904
2014	0.003744
2015	0.003521
2016	0.003424
2017	0.003456
2018	0.003392
2019	0.003328
2020	0.003072
2021	0.002911
2023	0.002753

Figure 23: Screenshot of table showing results of numerous calculations.

4.5 Results

The numerous regressions and calculations carried out throughout this project resulted in me finding the following final data for 2023.

Expected number of children born in 2023 = 596,489
Expected number of pregnancies in 2023 = 587,663
Expected % of pregnancies that were carrying multiple children = 1.4666%
Expected % of women who smoked during pregnancy in 2023 = 8.6%
Expected number of women who smoked during pregnancy in 2023 = 50,539
Expected number of children born via multiple carriage pregnancies in 2023 = 17,444
Expected number of children born victims of MSP in 2023 = 51,298
Expected number of children born with asthma in 2023 = 11,930
Expected proportion of all children born, whose mothers smoked during pregnancy and have asthma in 2023 = 0.002753
Expected proportion of all children born with asthma, whose mothers smoked during pregnancy in 2023 = 0.1376

5 Analysis

5.1 Testing Significance

The conclusion I wanted to make was to determine if my hypothesis that the number of women who smoked during pregnancy did directly affect the number of children born with asthma, was statistically significant. I decided to use a multiple linear regression for this.

The results of that regression can be found below;

Correlation matrix (pearson)			
	Y - expected number of children born with asthma	X ₁ - no. live births	X ₂ - % of women smoking during pregnancy
Y - expected number of children born with asthma	1	0.929814	0.929814
X ₁ - no. live births	0.929814	1	0.929814
X ₂ - % of women smoking during pregnancy	0.929814	0.929814	1

Figure 24: Correlation coefficient matrix between the numerous sets of inputs X_i and, observed data Y.

Coefficient Table Iteration 1 (adjusted R-squared = .1)								
	Coeff	SE	t-stat	lower t _{0.025} (R)	upper t _{0.025} (R)	Stand Coeff	p-value	VIF
b	4.715935	2.386815	1.975827	-0.788072	10.219941	0	0.0833867	
X ₁ - no. live births	0.0199813	0.00000659242	3030.950096	0.0199661	0.0199965	0.998969	-2.22045e-16	7.353716
X ₂ - % of women smoking during pregnancy	0.733931	0.218197	3.363617	0.230768	1.237095	0.00110861	0.00987864	7.353716

Figure 25: Coefficient table iteration, showing p value of each individual set of inputs on observed data Y.

$$Y - \text{expected number of children born with asthma} = 4.715935 + 0.0199813 X_1 - \text{no. live births} + 0.733931 X_2 - \% \text{ of women smoking during pregnancy}$$

Figure 26: Calculated equation for Y given two inputs for the two sets of inputs (X_1 & X_2).

The results obtained confirmed that both the number of live births each year and the percentage of women who smoked during each year, had immensely strong statistical significance in impacting the trend of the total number of children born with asthma each year.

The respected p value between X_1 and Y was of order -16, and hence was astonishingly close to zero, suggesting near perfect statistical significance. Furthermore, the correlation coefficient, R was found to be 1 between these two trends further reinforcing this statistical significance.

Likewise, the p value between X_2 and Y was 0.00987864, again an astonishingly small result which alongside the extremely high calculated correlation coefficient

R of 0.929674 suggested extraordinarily strong statistical significance also.

Given that one-tailed significance tests were carried out at the 5% significance level, a p value < 0.05 was required from both sets of inputs to determine statistical significance. Both sets of inputs were significantly below this threshold and hence within the critical region, there is clear affirmation that the proportion of women who smoked during pregnancy and the number of live births each year, were statistically significant in determining the expected number of children born with asthma each year.

6 Conclusion

In this Investigation I set out to determine how I could use statistical methods to determine if there was a relationship between MSP rates and levels of asthma In England and Wales, as well as make numerous predictions regarding the topics at hand for 2023. Furthermore, I wanted to refine and outline the modelling process, as well as specific mathematical techniques which acted as valuable tools when making predictions.

The PPDAC model is perhaps one of the best when undergoing any modelling process, as it allows for a very simple and coherent structure to be laid out, which is clear and easy to not only formulate, but also follow. The problem identified in mine was the same as the title of this project and through a series of planning, data collection, mathematical learning and analysis I was able to draw tangible results which were conclusive to the question and problem at hand.

As seen from section 4 through constant modelling and refining and from taking results and using them, I was able to come up with a series of relevant and meaningful predictions. The concluding results allowed me to formally predict, numerous figures for 2023, relating birth numbers, asthma rates and MSP rates all in one. The most significant perhaps being the number of children estimated to have asthma in 2023, the percentage of women undergoing MSP in 2023 and the proportion of all children who were victims of MSP who have asthma (around 1 in 9 or a proportion of 0.1376).

A key part of all investigations however is of course determining whether results are of significance. This was something that I was able to attain using correlation statistics such as p-values and regression coefficients etc. Furthermore, I was able to take my results and prior data to establish confirmation that there is a positive correlation between MSP rates and asthma levels in England and Wales using a multiple linear regression and correlation coefficients as well as p-values. This was hence a strong part of my conclusion as I was able to statistically prove that there is a direct impact of MSP on asthma rates in the UK.

In conclusion, the results of this investigation allowed me to determine quantitative predictions for numerous different measures, enabling me to reach the aim of using statistical methodology to predict the impact of MSP on asthma rates as well as simply just figures involving MSP, childbirth rates and asthma levels in the nation. Furthermore, I was able to conclusively and empirically prove a relationship between levels of MSP and asthma rates in children. Thus, the main aims of this investigation had been met.

7 Personal Evaluation & Next Steps

Throughout this project I learnt an awful lot about myself and developed many skills such as research abilities, time management, organisation, and my mathematical skill set, all whilst gaining experiencing in writing in a formal, academic manner.

I displayed good evidence of all these skills and more. However, a key area in which I had developed significantly throughout was my time management and planning. I never rushed my project and always managed to meet deadlines set, however in my original plan set out and my Gantt chart created, I did not anticipate the immense workload I would endure throughout the whole project regarding my studies and other commitments outside of the EPQ project. If I were to redo this project, I would try better to account for this workload and plan the stages of my project in a different manner.

I also found throughout the project that researching and sourcing data was a much more tedious task than originally anticipated. I also found problems regarding trying to combine all my findings and data and this often took prolonged periods of thinking and planning to combine it in a way that made sense and was coherent and useful.

My key-dilemma throughout my project was a lack of sense of direction. Despite having goals in mind, it took me an exceedingly long while to determine exactly where my project was to be headed at many stages throughout the project, specifically throughout the modelling stage. However, as time went on, I realigned myself in the correct direction, making sure that I was stopping and considering the usefulness of every stage that I carried out, towards my project.

My research was very thorough. Using a reference class of 10 years was also helpful, as it not only reduced the impact that anomalous results would have, but it also allowed me to observe reliable trends over time, which gained more integrity, with the more data gathered. Further to this, the method of forecasting I used throughout was reference class forecasting, a method which revolves around the accumulation of past data, so having data from up to 10 years ago proved to be especially useful for my project.

Heading into this project, the outcome I had in mind was to establish estimates for the rate of Asthma in England and Wales in 2023, as well as determining a relationship between smoking during pregnancy and the rates of asthma. I have found the answers to all questions I had, providing numerous quantitative results regarding the levels of asthma in this region in 2023, as well as drawing the conclusion that reducing the rate of smoking during pregnancy, has a direct impact on the levels of asthma, and was able to prove this by establishing statistical significance between the two trends.

If I were to repeat the project, I would most definitely organise my gantt chart and plan, to allow for more time in certain stages of my project. Specifically on the research side as I already discussed how this was something that I spent a lot longer on than I originally suspected. I would also like to increase the years from which I secured data as this would only enable me to draw more accurate and reliable trends and hence results. However, the key change that I would make, is that I would make sure from early on, that I knew exactly where I would be headed as I worked on through the project, and exactly what I will be hoping for in the final write up of my project. This is certainly something I would be able to incorporate next time as I believe a huge reason as to why I experienced this difficulty was due to lack of experience in projects such as this, however I have now learnt from this, and have been able to better my ability in working on such tasks.

8 Citations

[1] – changing minds (no date) The Need to Predict, The need to predict. Available at: <http://changingminds.org/explanations/needs/prediction.htm#:text=One%20of%20the%20things%20we%20are%20constantly%20doing,us%20a%20lot%20better%20chance%20to%20control%20things%E2%AF> (Accessed: October 5, 2022).

[2] – Vara, V. (2017) ”we will literally predict their life outcomes” — backchannel, Wired. Conde Nast. Available at: <https://www.wired.com/2016/05/we-will-literally-predict-their-life-outcomes/> (Accessed: November 2, 2022).

[3] – NCT (National Childbirth Trust) (2019) Smoking during pregnancy: Pregnancy articles support: NCT, NCT (National Childbirth Trust). Available at: <https://www.nct.org.uk/pregnancy/food-and-nutrition/smoking-during-pregnancy#:text=Babies%20and%20children%20whose%20mothers%20smoke%20during%20pregnancy,as%20adverse%20behaviour%203%20performing%20poorly%20at%20school.> (Accessed: November 24, 2022).

[4] – Zacharasiewicz, A. (2016) Maternal smoking in pregnancy and its influence on childhood asthma, ERJ open research. U.S. National Library of Medicine. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5034599/> (Accessed: November 25, 2022).

[5] – Frost, J. (2021) Making predictions with regression analysis, Statistics By Jim. Available at: <https://statisticsbyjim.com/regression/predictions-regression/:text=You%20can%20use%20regression%20equations%20to%20make%20predictions.,between%20each%20independent%20variable%20and%20the%20dependent%20variable.> (Accessed: November 2, 2022).

[6] – Spiegelhalter, D. (2019) The Art of Statistics: Learning From Data. London: Pelican, an imprint of Penguin Books.

[7] – Wittwer, J. (2004) Articles, Deterministic model example. Available at: <https://www.vertex42.com/ExcelArticles/mc/DeterministicModel.html#:text=A%20deterministic%20model%20is%20a%20model%20that%20gives,monthly.%20The%20model%20is%20just%20the%20equation%20below%3A> (Accessed: November 19, 2022).

[8] – Kenton, W. (2022) Why stochastic modeling is less complicated than it sounds, Investopedia. Investopedia. Available at: <https://www.investopedia.com/terms/s/stochastic-modeling.asp> (Accessed: November 19, 2022).

[9] – MacKay, R.J. and Oldford, R.W. (2000) Scientific method, statistical method and the speed of light, Project Euclid. Institute of Mathematical Statistics. Available at: <https://projecteuclid.org/journals/statistical-science/volume->

15/issue-3/Scientific-Method-Statistical-Method-and-the-Speed-of-Light/10.1214/ss/1009212817.full (Accessed: November 19, 2022).

[10] - Sindhuja, S. (2017) Mathematical models: Types, structure and advantages: Decision making, Essays, Research Papers and Articles on Business Management. Available at: <https://www.businessmanagementideas.com/management/decision-making-management/mathematical-models-types-structure-and-advantages-decision-making/10034> (Accessed: November 19, 2022).

[11] - IBM (no date) About linear regression, IBM. Available at: <https://www.ibm.com/topics/linear-regression#:text=Why%20linear%20regression%20is%20important%20Linear-regression%20models%20are,to%20various%20areas%20in%20business%20and%20academic%20study.> (Accessed: November 20, 2022).

[12] - Conceptually (no date) Reference class forecasting - definition and examples, Conceptually. Available at: <https://conceptually.org/concepts/reference-class-forecasting> (Accessed: January 22, 2023).

[13] - JMP (no date) Multiple linear regression, JMP. Available at: https://www.jmp.com/en_k/statistics-knowledge-portal/what-is-multiple-regression.html (Accessed : November 20, 2022).

[14] - Berk, M. (2022) How does linear regression really work?, Medium. Towards Data Science. Available at: <https://towardsdatascience.com/how-does-linear-regression-really-work-2387d0f11e8> (Accessed: November 20, 2022).

[15] - Indeed Editorial Team (2021) Prediction interval vs. confidence interval: Differences and ... - indeed, Prediction Interval vs. Confidence Interval: Differences and Examples. Available at: <https://www.indeed.com/career-advice/career-development/prediction-interval-vs-confidence-interval> (Accessed: March 4, 2023).

[16] - Frost, J. (2018) Why are there no P values in nonlinear regression?, Statistics By Jim. Available at: <https://statisticsbyjim.com/regression/no-p-values-nonlinear-regression/:text=P%20values%20for%20the%20independent%20variables%20in%20linear,controlling%20for%20the%20other%20variables%20in%20the%20model.> (Accessed: January 22, 2023).

[17] - Kumar, A. (2022) Interpreting F-statistics in linear regression: Formula, examples, Data Analytics. Available at: <https://vitalflux.com/interpreting-f-statistics-in-linear-regression-formula-examples/> (Accessed: January 22, 2023).

[18] - Bhandari, P. (2022) Correlation coefficient: Types, formulas examples, Scribbr. Available at: <https://www.scribbr.com/statistics/correlation-coefficient/> (Accessed: January 22, 2023).

[19] - Synowiec, P. (2021) Births in England and Wales: 2020, Births in England and Wales - Office for National Statistics. Office for National Statistics. Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthsummarytablesenglandandwales/2020> (Accessed: January 22, 2023).

[20] - Office for National Statistics (2021) Births in England and Wales Statistical Bulletins, Births in England and Wales Statistical bulletins - Office for National Statistics. Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthsummarytable-senglandandwales/previousReleases> (Accessed: January 22, 2023).

[21] - Centers for Disease Control and Prevention (2022) Most recent national asthma data, Centers for Disease Control and Prevention. Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/asthma/most_recent_national_asthma_data.htm (Accessed: January 22, 2023).

[22] - Stewart, C. (2022) Smoking during pregnancy in England 2006-2022, Statista. Available at: <https://www.statista.com/statistics/445149/smoking-during-pregnant-in-england/> (Accessed: January 22, 2023).

[23] - Michas, F. (2022) Multiple birth rate in England and Wales, Statista. Available at: <https://www.statista.com/statistics/971604/england-and-wales-multiple-birth-rate/> (Accessed: January 22, 2023).

[24] - American Society for Reproductive Medicine (2012) Multiple pregnancy and birth: Twins, triplets, and high order Multiples (booklet), Multiple Pregnancy and Birth: Twins, Triplets, and High Order Multiples (booklet). Available at: <https://www.reproductivefacts.org/news-and-publications/patient-fact-sheets-and-booklets/documents/fact-sheets-and-info-booklets/multiple-pregnancy-and-birth-twins-triplets-and-high-order-multiples-booklet/:.text=Naturally%2C%20twins%20occur%20in%20about%20one%20in%20250,of%20infertility%20treatment%2C%20but%20there%20are%20other%20factors>. (Accessed: January 22, 2023).

[25] - Hitti, M. (2005) Smoking while pregnant ups baby's asthma risk, Smoking While Pregnant Ups Baby's Asthma Risk. WebMD. Available at: <https://www.webmd.com/baby/news/20050411/smoking-while-pregnant-ups-babys-asthma-risk:.text=Children%20whose%20mothers%20smoked%20for%20any%20part%20of,who%20had%20ever%20smoked%20quit%20before%20getting%20pregnant>. (Accessed: January 22, 2023).