Predicting Life Expectancy in MSOAs – Building a Linear Regression Model.

Building The Model (Phase 1)

<u>Model Choice</u>

A linear regression model was chosen to predict life expectancy for several reasons. From conducting exploratory data analysis, it was evident that the predictor variables had a linear relationship with life expectancy (Figure 1), making linear regression a suitable model for the data. Secondly, a linear regression model can predict the value of an outcome based on multiple predictor variables, making it well-suited to incorporate the multifaceted determinants of health. Furthermore, according to Schneider et al. (2010), linear regression should be utilised when the dependent variable is continuous. Since life expectancy is indeed continuous, a linear regression model was thought to be a suitable choice. These reasons demonstrate the appropriateness of a linear regression model to predict life expectancy.

<u>Predictor Variables:</u>

The following predictor variables were chosen: net annual income before household costs; Index of Multiple Deprivation (IMD) scores, and economic inactivity due to long-term sickness/disability. All predictors were related to Middle Layer Superior Output Areas (MSOAs).

To prevent multicollinearity, the decision was made to include only one income-based statistic. Net annual income before housing costs (equivalised) was included to focus on overall income without separating housing from daily living expenses. Equivalised data was chosen to enable more reliable comparisons across different regions. This distribution is illustrated in Figure 2. IMD scores were sourced from mySociety, a non-for-profit social enterprise. The scores were included as Charlton et al. (2013) explain that deprivation is correlated with higher incidences of morbidity and mortality, making it an important factor to consider in life expectancy. Similarly, economic inactivity due to long-term sickness/disability, sourced from NOMIS, was incorporated to explore the impact of health inequalities on life expectancy. Including multiple determinants in the model allows for more nuanced policy recommendations – this is important because addressing economic inactivity due to health issues may require different interventions than addressing general deprivation, for example.

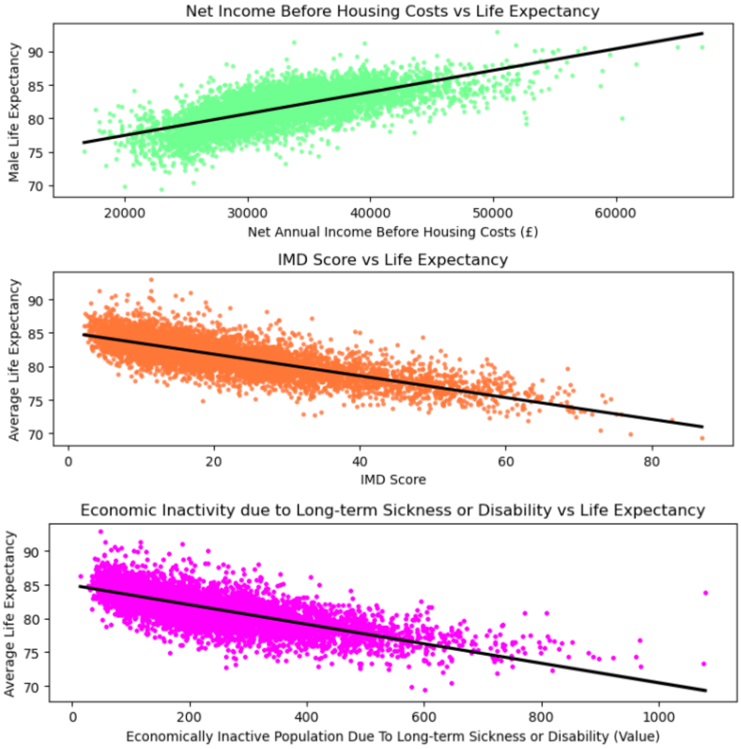


Figure 1. Scatter plots with regression lines showing the relationship between the predictor variables and life expectancy.

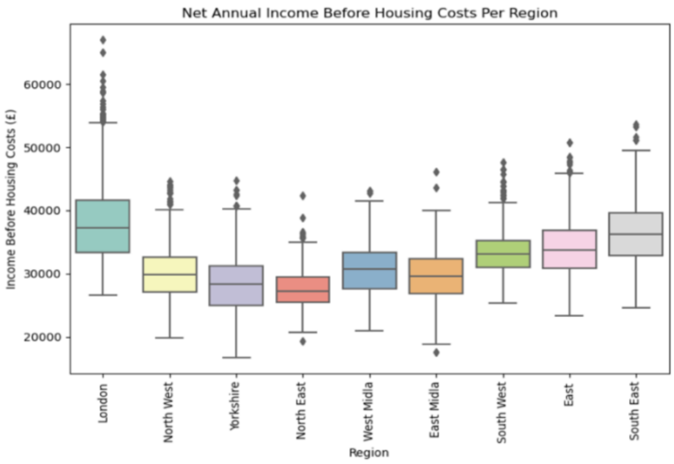
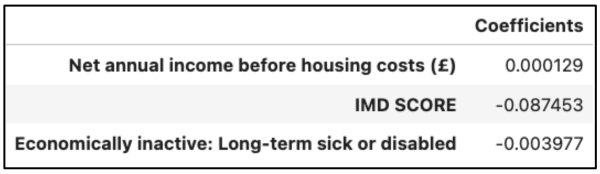
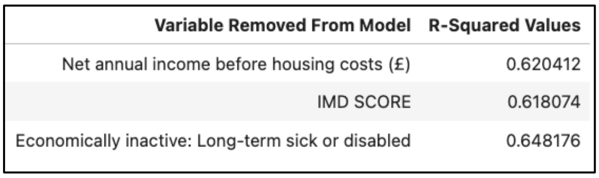


Figure 2. A box plot showing the distribution of net annual income before housing costs per region.



	Coefficients
Net annual income before housing costs (£)	0.000129
IMD SCORE	-0.087453
Economically inactive: Long-term sick or disabled	-0.003977

Table 1. Regression coefficients for the different predictors.



Variable Removed From Model	R-Squared Values
Net annual income before housing costs (£)	0.620412
IMD SCORE	0.618074
Economically inactive: Long-term sick or disabled	0.648176

Table 2. Comparing the impact of the variables on life expectancy using R^2.

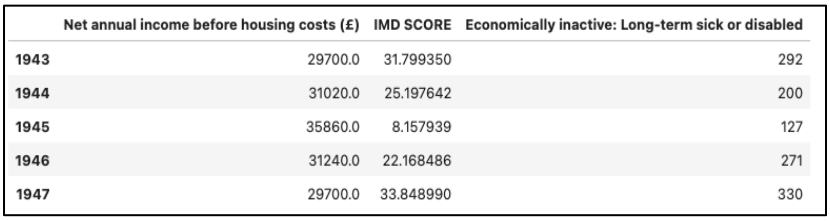
<u>Impact of Predictors on Life Expectancy:</u>

The regression coefficients quantify the relationships between the predictor variables and life expectancy (Table 1). As the net annual income before housing cost increases by £1, life expectancy is predicted to increase by 0.000129 years (0.047 days). Moreover, as the IMD score increases by 1 (indicating higher deprivation), life expectancy is predicted to decrease by 0.087453 years (31.94 days). This suggests that greater deprivation is associated with shorter life expectancy. Likewise, for each additional person who is economically inactive due to long-term sickness or disability, life expectancy is predicted to decrease by 0.003977 years (1.452 days). To find which of these variables had the most significant impact on life expectancy, each variable was removed one-by-one from the data and the model refit. The R^2 was then calculated for each model to identify which of the three yielded the greatest reduction in R^2 relative to the original model. As shown in Table 2, removing 'IMD Score' resulted in the model with lowest R^2, thus this predictor has the most significant impact on life expectancy.

Predicting Life Expectancy (Phase 2)

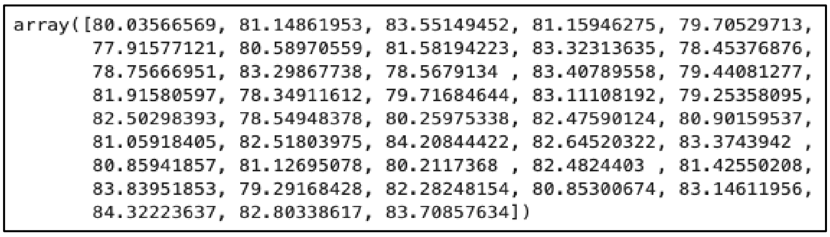
Calculating Increased Life Expectancy in Dudley:

1. The dataframe (final_df) was filtered to select for rows where the MSOA column contained 'Dudley'. The filtered data was assigned to a new dataframe called dudley_df.
2. From dudley_df, the three predictor value columns ('Net annual income before housing costs (£)', 'IMD SCORE' and 'Economically inactive: Long-term sick or disabled') were filtered for. The filtered data was assigned to a new dataframe called mini_dudley_df.
3. From mini_dudley_df, the values in the 'Net annual income before housing costs (£)' column were multiplied by 1.1 to account for the 10% increase in income (Table 3).
4. mini_dudley_df was passed into the model to predict the change in life expectancy (Table 4).



	Net annual income before housing costs (£)	IMD SCORE	Economically inactive: Long-term sick or disabled
1943	29700.0	31.799350	292
1944	31020.0	25.197642	200
1945	35860.0	8.157939	127
1946	31240.0	22.168486	271
1947	29700.0	33.848990	330

Table 3. mini_dudley_df dataframe, The values in 'net annual income before housing costs' have been multiplied by 1.1.



```
array([80.03566569, 81.14861953, 83.55149452, 81.15946275, 79.70529713,
       77.91577121, 80.58970559, 81.58194223, 83.32313635, 78.45376876,
       78.75666951, 83.29867738, 78.5679134 , 83.40789558, 79.44081277,
       81.91580597, 78.34911612, 79.71684644, 83.11108192, 79.25358095,
       82.50298393, 78.54948378, 80.25975338, 82.47590124, 80.90159537,
       81.05918405, 82.51803975, 84.20844422, 82.64520322, 83.3743942 ,
       80.85941857, 81.12695078, 80.2117368 , 82.4824403 , 81.42550208,
       83.83951853, 79.29168428, 82.28248154, 80.85300674, 83.14611956,
       84.32223637, 82.80338617, 83.70857634])
```

Table 4. Predicted life expectancy in Dudley following a 10% increase in net annual income before housing costs.

Reliability of Findings:

The model was reliably able to predict life expectancy in Dudley for various reasons. Firstly, the data utilised to build the model was of high-quality, having been obtained from trustworthy sources. As each data source covered MSOAs, the model could appropriately predict life expectancy in Dudley MSOAs. Secondly, the scatter plot of the predicted vs actual values (Figure 3) shows a 45-degree regression line with points tightly clustered around the line, indicating strong model performance. The normally distributed residuals (Figure 4) also suggests that the model suitably captures the relationship between the predictor variables and life expectancy. These evaluation methods show the model is reliably able to make predictions. Lastly, the R-squared value is 0.658 (to 3 decimal places) which implies that, while the model may be less precise than one with a higher R-squared value, it is sufficiently useful to predict life expectancy (Hamilton et al. 2015). These factors explain the reliability of the model to predict life expectancy in Dudley.

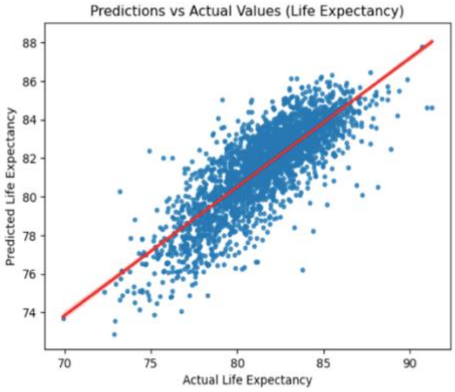


Figure 3. Scatter plot showing the predictions vs actual life expectancy.

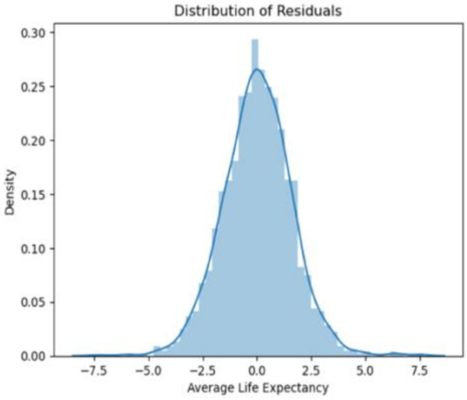


Figure 4. Residuals are normally distributed, indicating that linear regression is a suitable model choice for the data.

However, life expectancy might not increase as predicted by the model for a few reasons. Firstly, determinants of life expectancy are multifactorial. Whilst the model includes three predictor variables, it does not account for factors such as pollution levels, barriers to health and lifestyle choices which also play a significant role in life expectancy. Additionally, public policies or new health interventions could alter the relationship between predictors and life expectancy. For example, Gredner et al. (2021) estimated that stricter tobacco control policies in Europe could prevent 1.65 million lung cancer cases over two decades. If such policies were introduced, they could potentially increase life expectancy for smokers more effectively than an increase in income. Lastly, the positive impact of increased income on life expectancy may diminish beyond a certain point, leading to a plateau. These factors collectively explain why life expectancy in Dudley might not increase as the model predicts.

Bibliography

Charlton, J. et al. 2013. Impact of deprivation on occurrence, outcomes and health care costs of people with multiple morbidity. *Journal of Health Services Research & Policy* 18(4), pp. 215–223. doi: 10.1177/1355819613493772.

Gredner, T. et al. 2021. Impact of tobacco control policies implementation on future lung cancer incidence in Europe: An international, population-based modeling study. *The Lancet Regional Health - Europe* 4, p. 100074. doi: 10.1016/j.lanepe.2021.100074.

Hamilton, D.F. et al. 2015. Interpreting regression models in clinical outcome studies. *Bone & Joint Research* 4(9), pp. 152–153. doi: 10.1302/2046-3758.49.2000571.

Schneider, A. et al. 2010. Linear Regression Analysis. *Deutsches Ärzteblatt International* 107(44), pp. 776–782. doi: 10.3238/arztebl.2010.0776.