# Does higher daily caloric intake leads to lower life expectancy?

Ilja Savčenko, Danielius Michailovas, and Rimgaudas Jurgaitis

ISM University of Management and Economics

ECO 105: Econometrics

PhD. Aleksandr Christenko

Jan. 10, 2025

## Hypothesis

Over the last few decades, the question of proper diet and right amount of food consumption has been on the rise. Concepts of calorie counting and “superfoods” started spreading, and fears of pesticides and additives in what we eat fuelled the organic food movement. Of course, as obesity rate keeps growing worldwide, it proves that simple trends and health concerns are not enough to soothe the human appetite (Ritchie & Roser, 2024). One of the biggest contributors to this problem is urbanization. Centralization of work and food production drastically raised the availability of food. Furthermore, the rapid development of food industry, which, with its marketing ploys and borderline addictive products, have drastically increased the consumption. However, those weren’t always the main diet related issues.

In the past undereating was caused mainly by two reasons – food deficit and terrible work conditions. Jobs were more manual and labor-intensive which did not leave enough time for proper nutrition. This would lead to health complications and early deaths either from starvation or improper diet. From the late 20th century food security became less of a concern since the Green Revolution in the 1940s – 1970s increased food availability (Pingali, 2012). Moreover, automatization and technological advancements made work easier and more efficient. On the other hand, one problem was replaced by another – people started living sedentary lifestyles and overeating. Consequently, obesity became a global issue with more people dying at a younger age. A question of how daily caloric intake affects life expectancy arises. The aim of this research essay is to test whether higher daily caloric intake leads to lower life expectancy using time-series regression analysis.

## Data commentary

The data for the time-series regression analysis was taken from Our World in Data (OWID). Our World in Data (OWID) is a scientific online publication that focuses on large global problems. Collected data about various factors that could influence life expectancy of United Kingdom citizens during 1842 - 2023 period was combined into a dataset.

## Merging columns

Original dataset for *gdp* and *relative_health_expenditure* has a few missing values in the current years. Since there are data sources that do cover these missing years, but with different measurement methodology, it was decided that merging data would lead to more accurate world representation than imputation of the values. To add missing values in original column, this formula was used:

**Formula 1.**

$$a_t = (k * \frac{n_t - n_{t-1}}{n_{t-1}} + 1) * a_{t-1},\ n - \ other\ data;\ k - \ adjustment\ factor,\ a\ - original\ data$$

Adjustment factor was calculated with this formula:

**Formula 2.**

$$k = \frac{\sum_{t=2}^{t} \frac{\frac{a_t - a_{t-1}}{a_{t-1}}}{\frac{n_t - n_{t-1}}{n_{t-1}}},}{t-1}, a - \ original\ data;\ n - \ other\ data;\ t - \ time\ period$$

Adjustment factor ought to reflect the difference in rate of change for the same years. In reality, for both column merges, if was hovering around 1 (1.01 for *gdp* and 1.02 for *relative_health_expenditure*) . There might be some issues regarding adjustment factor being calculated for the whole range and not around the recent values.

## Imputations

Our time frame is spanning to the year of 1842, so it is natural that there are missing values or values are recorded infrequently (each 10 years). To be particular, *urbanization_rate*, *daily_caloric_intake*, *average_working_hours* columns have less frequent measurements in the past and a few missing values in modern days. Column of *lighting_price* also has a few missing values in the modern days and it was decided not to merge this one, because no similar datasets were found, so we chose imputation.

For the imputation process itself, an iterative imputer from scikit-learn library in python was chosen with an estimator of random forest, rather than the default of Bayesian ridge. The dataset consists mostly of exponential variables rather than linear ones, so the nature of decision trees in random forest approach should lead to better results, compared to Bayesian ridge. Iterative imputer was chosen because of the same reasons mentioned earlier. To get as close to the logical sequence of the trend, default arguments of iterative imputer were modified as follow:

- Maximum iterations increased from 10 to 100.
- Minimum possible imputed value set to 0.
- 3 nearest values are taken into consideration, instead of 0.

Imputed values are close enough to what someone could expect from the trend continuation, so it was decided to keep the imputation process unchanged. There are a few imperfections, including convergence to a single value or jumps around actual values, but this should not lead to any significant impact on the analysis.

Variables used in the analysis are presented in Table 1 below.

**Table 1.**

*Variable description.*

| **Variable** | **Explanation** |
|---|---|
| Year | Time variable representing annual observations (from 1842 to 2023). |
| Life expectancy | Dependent numeric variable that suggests the period life expectancy at birth in years. |
| GDP | Independent numeric variable that determines the Gross domestic product (GDP) expressed in British pounds, adjusted for inflation. |
| Relative health expenditure | Independent numeric variable that captures the spending on government funded health care systems and social health insurance, as well as compulsory health insurance as a share of GDP (in %). |
| Urbanization rate | Independent numeric variable that shows share of the population living in urbanized areas (in %). |
| Daily caloric intake | Independent numeric variable that measures the supply of calories per person per day. |
| Average working hours | Independent numeric variable that describes the average annual working hours per worker. |
| Child mortality | Independent numeric variable that shows the estimated share of newborns who die before reaching the age of five (in %). |
| World war 1 | WW1 dummy variable indicating the war period (with a 1) and period before and after the war (with a 0). |
| World war 2 | WW2 dummy variable indicating the war period (with a 1) and period before and after the war (with a 0). |
| Lighting price | Independent numeric variable that represents the price of lighting per million lumen-hours in British Pound. |

## Describing data

The descriptive statistics below give us a clear picture of the *life_expectancy* variable in our dataset. The average life expectancy is 59.927, which is very close to the median value of 60.675. This shows that the data is mostly symmetrical. The skewness is -0.02684, which means there is a very slight tendency for the data to have a few lower values, but this effect is minimal.

**Table 2.**

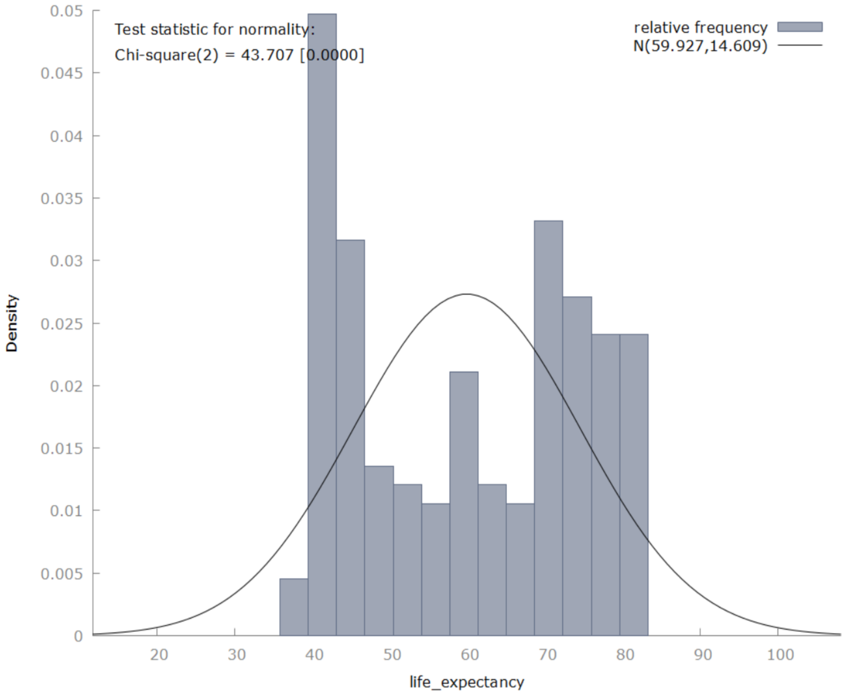
*Summary of the dependent variable descriptive statistics.*

| **Statistic** | **Value** |
|---|---|
| Median | 59.927 |
| Median | 60.675 |
| Minimum | 37.680 |
| Maximum | 81.440 |
| Standard deviation | 14.609 |
| Skewness | -0.026864 |
| Ex. kurtosis | -1.5487 |
| Missing obs. | 0 |

If we look from the visual perspective using a simple frequency distribution, we can see that the *life_expectancy* variable deviates from normality (see Figure 1). The data shows a slight negative skewness with a longer left tail.

**Figure 1.**

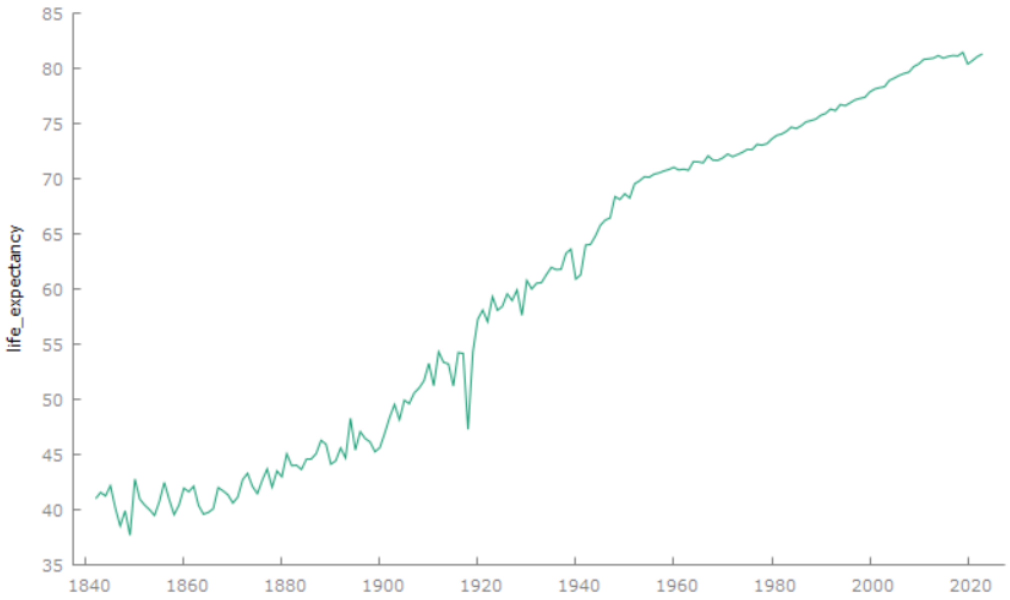
*Frequency of “life_expectancy” variable.*



Looking at Figure 2, time series plot of *life_expectancy* we can see an obvious upwards trend, which was expected due to the tendency of human life span increasing as technologies progress. There was no observed seasonality in the plot, however there are a couple of noticeable outliers (in 1918 drop by 6.88 and in 1940 drop by 2.71) which can be attributed to WW1 and WW2. Overall, based on Table 2 and Figure 2 we can clearly tell see increasing mean and non-constant variance, which implies presence of a unit root.

**Figure 2.**

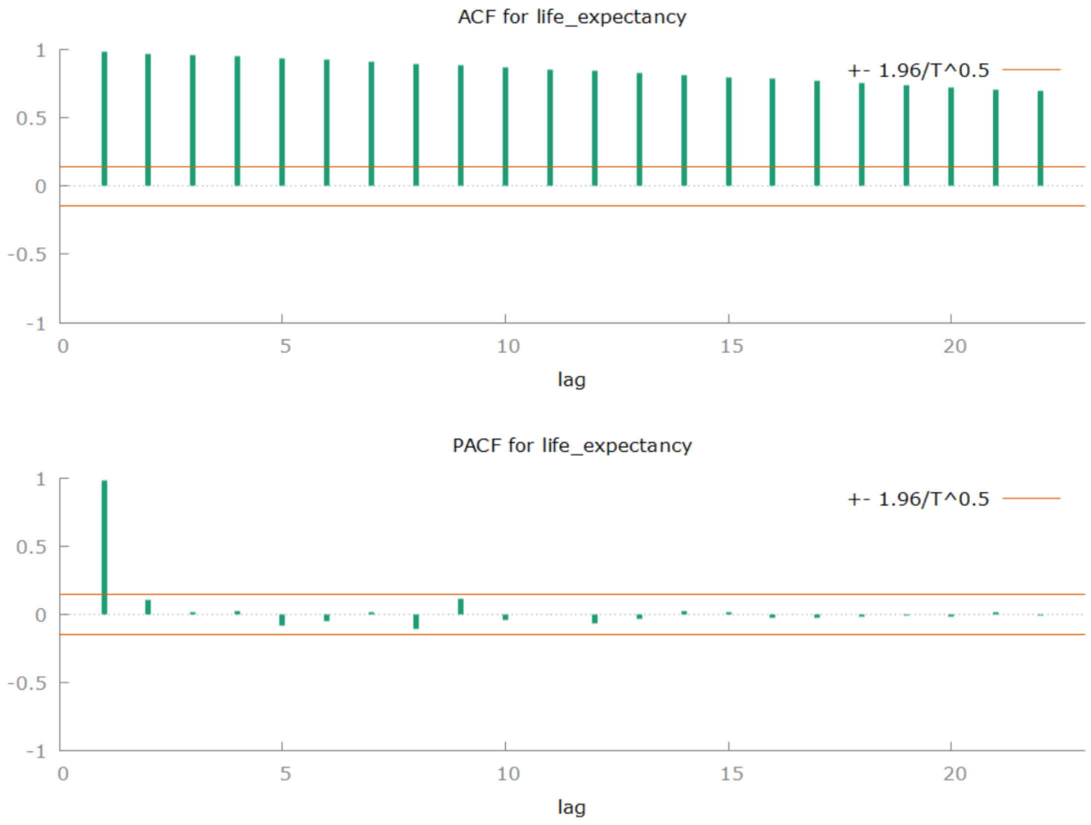
Time series plot of “life_expectancy”.



In Figure 3, we focus on the ACF for *life_expectancy*. The Autocorrelation Function for our dependant variable shows strong autocorrelation at all lags, indicating non-stationarity. However, the presence of strong autocorrelation at smaller lags indicates some persistence in the data.

**Figure 3.**

*Autocorrelation Function plot of "life_expectancy" variable.*



## Stationarity test

As noted before, when checking time series plot, both ADF (p-value of 0.33) and KPSS (p-value < 0.01) results indicate non-stationarity of *life_expectancy*. This implies both non-constant mean and variance in our dependant variable. To fix that, we transform *life_expectancy* via first difference.

After transformation, the results from ADF (p-value of 0.0001) and KPSS (p-value of 0.076) rejected / didn't reject the null hypothesis, respectively. This proves that after transformation of *life_expactancy*'s (now *d_life_expactancy*'s) mean and variance are constant, and should produce optimal results.

**Table 3.**

*Augmented Dickey-Fuller test and Kwiatkowski–Phillips–Schmidt–Shin test for "relative_health_expenditure", "urbanization_rate", "daily_caloric_intake" variables*

| | **No Difference** | | **First Difference** | | **Percentage Difference** | |
|---|---|---|---|---|---|---|
| Stationarity tests | ADF | KPSS | ADF | KPSS | ADF | KPSS |
| relative_health_expenditure | 1 | < .01 | 0.0005418 | > .10 | 0.003182 | < .01 |
| urbanization_rate | 0.6584 | < .01 | 0.05181 | 0.085 | 0.2133 | > 0.1 |
| daily_caloric_intake | 0.828 | < .01 | 1.891e-61 | > .10 | 5.174e-62 | > 0.1 |

All other numeric variables also suffered from unit root, so we analysed them for proper difference transformations. *relative_health_expenditure*, *urbanization_rate*, *daily_caloric_intake* become stationary with first difference transformation, see Table 3. *urbanization_rate*'s ADF results (p-value of 0.05181) may be above the threshold of 0.05, but we judged it to be close enough to still use first difference.

**Table 4.**

*Augmented Dickey-Fuller test and Kwiatkowski–Phillips–Schmidt–Shin test for "gdp" variable*

| | **No Difference** | | **First Difference** | | **Percentage Difference** | |
|---|---|---|---|---|---|---|
| Stationarity tests | ADF | KPSS | ADF | KPSS | ADF | KPSS |
| relative_health_expenditureelative_health_expenditure | 1 | < .01 | 0.0673 | 0.044 | 0.01002 | > .10 |

The results from numerical value independent variable *gdp*, however, show that first difference does not solve non-stationarity in means. That is why we opted for percentage difference transformation, who's p-values indicate lack of unit root.

**Table 5.**

*Augmented Dickey-Fuller test and Kwiatkowski–Phillips–Schmidt–Shin test for "average_working_hours", "child_mortality", "lighting_price" variables.*

| | **No Difference** | | **First Difference** | | **Percentage Difference** | | **Double First Difference** | |
|---|---|---|---|---|---|---|---|---|
| Stationarity tests | ADF | KPSS | ADF | KPSS | ADF | KPSS | ADF | KPSS |
| average_working_hours | 0.9106 | < .01 | 0.1125 | < .01 | 0.1317 | 0.013 | 0.004354 | > .10 |
| child_mortality | 0.3813 | < .01 | 0.0002464 | 0.028 | 0.0861 | 0.018 | 6.814e-28 | > .10 |
| lighting_price | 0.004839 | < .01 | 5.399e-30 | < .01 | 0.05568 | 0.036 | 1.014e-10 | > .10 |

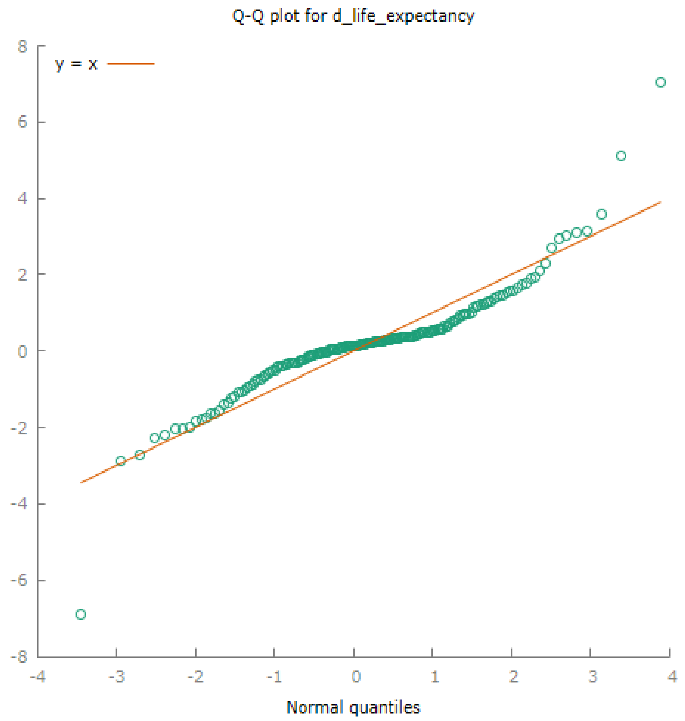
Lastly, we couldn’t solve non-stationarity of *average_working_hours*, *child_mortality*, *ligting_price* variables. As we lack better difference transformation and for a better interpretation of results, we used double first difference transformation. Its effect on stationarity test results indicated at a constant mean and variance, see Table 5.

## Dependent variable analysis

Dependant variable *life_expectancy* with a taken first difference was tested for normality via the Doornik Hansen, Shapiro-Wilk, Lilliefors, Jarque-Bera tests, and Q-Q plot (see Figure 4). The results for all of the normality tests for *d_life_expectancy* show a p-value of 0 or approaching it, which implies that the tests have detected a significant difference between the data distribution and a normal distribution. The results of Doornik-Hansen test and Jarque-Bera test highlight that the dependent variable is likely not normally distributed in terms of skewness and/or kurtosis.

**Figure 4.**

*Q-Q plot for "d_life_expectancy".*



The results of the normality tests are further supported by the Q-Q plot, which depicts that quantiles of *d_life_expectancy* do not align with quantiles of a normal distribution. The majority of dependent variable observations are relatively near 0, however, the outliers make the curve positively skewed.

Overall, the test and plot results indicate that a fix for the non-normality of our dependent variable was needed. However, logarithmic transformation or other transformative actions to solve this problem were not taken since they would overcomplicate the interpretation of the results.

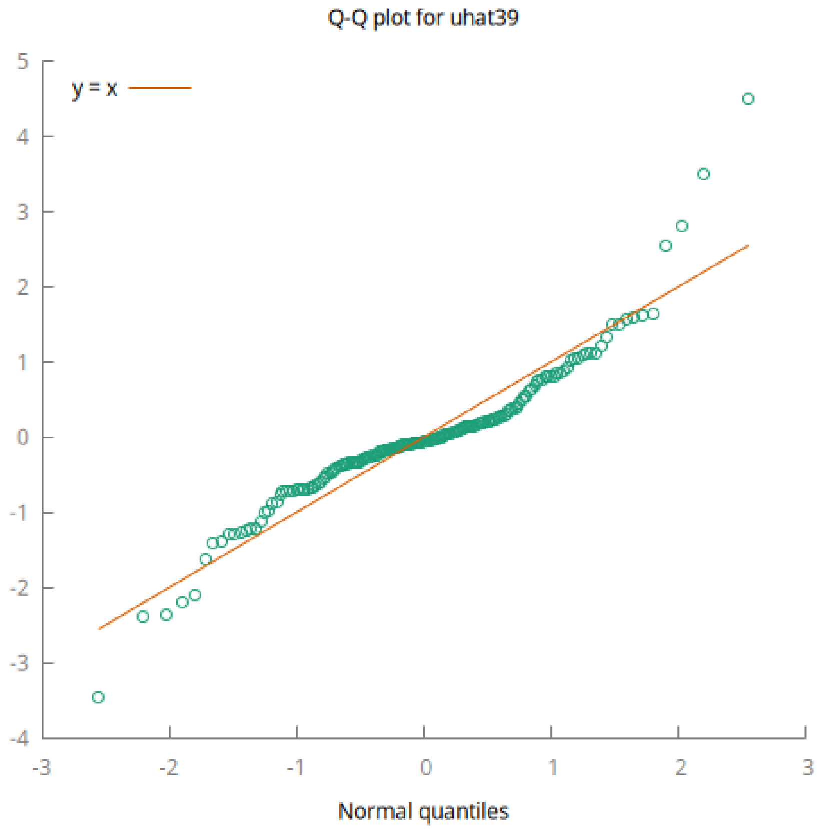
## Model

### Normality of residuals

Distribution of residuals is pretty normal if we look at the whole picture. There are 5 years, where the change in life expectancy was big, compared to other years. This makes 5 distinct outliers, which make the distribution not normally distributed (all tests agree on that). We examined these outliers and found no fundamental cause, which could be accounted for in our model. This is a minor issue, which already removes the possibility of regression being the best unbiased estimator and causes bias in p-values.

**Figure 5.**

*Q-Q plot for residuals.*



### Heteroscedasticity

Both White’s (p-value of 8.5358e-05) and Breush-Pagan (p-value of 3.97085e-12) indicate severe problem with heteroscedasticity of residuals. The problem can not be fixed, since there is no available data to increase length of the dataset or to add relevant variables that would span long enough. To account for this, we will use standard error, although the model can no longer be the best linear unbiased estimator.

### Multicollinearity

Our numeric variables have at least 1 first difference, so there should not be any spurious relationships. The table below shows exactly that, where all variables are perfectly fine and do not have any problems.

**Table 6.**

*Multicollinearity test results.*

| Variable | VIF |
|---|---|
| d_d_average_working_hours_3 | 1.089 |
| d_d_lighting_price_1 | 1.021 |
| d_relative_health_expenditu_5 | 1.045 |
| d_urbanization_rate_5 | 1.041 |
| d_daily_caloric_intake_2 | 1.007 |
| perc_gdp | 1.139 |
| ww1 | 1.014 |
| ww2 | 1.055 |
| d_d_child_mortality | 1.024 |

### Autocorrelation

Durbin-Watson test indicates a positive autocorrelation (p-value of 0.94 for positive autocorrelation). This is likely to be a sign of important missing variables (possibly: income inequality, air pollution, medicine innovations, child labor participation, etc.). Since there was no relevant data that is spanning long enough for our model, the model will not be updated.

To account for heteroscedasticity and autocorrelation issues, standard error (HAC) is being applied.

## Pseudo-causality

As mentioned earlier, both *d_life_expectancy* and *d_daily_caloric_intake* are stationary. All 3 of the criteria show different ideal lags (AIC - 3, BIC - 1, HQC - 2). To select the best lag out of these 3, a model for each lag ought to be built and evaluated. Clearly, 2 or 3 lag models produce better results (in Granger's causality), when comparing test results of autocorrelation, heteroscedasticity and normality of residuals. Difference between models of 2 and 3 lags is very minuscule and way better than with 1, so further analysis will be continued with both. In both models, no lags of opposite variables are able to reach significance level. F-tests show the same - there is no pseudo-causal relationship between life expectancy and daily caloric intake.

Results should be double checked using the Toda-Yamamoto approach. VAR lag selection procedure this time suggests using 2 (AIC) or 4 (BIC, HQC) lags. Since we need additional lag for fixing non-normality, models with 3 and 5 lags ought to be created. Results are as follows:

**Table 7.**

| **Dependent** | **Independent** | **p-value** |
|---|---|---|
| daily_caloric_intake | All lags of life_expectancy | 0.7207 |
| life_expectancy | All lags of daily_caloric_intake | 0.8104 |

Overall, pseudo-causal relationship is not detected when using either classical or Todo-Yamamoto approaches.

## Model interpretations

**Table 8.**

*Model results.*

| **Variable** | **Coefficient** | **p-value** |
|---|---|---|
| const | **0.20274** | **0.0002** |
| d_d_average_working_hours_3 | **−0.00575536** | **0.049** |

| | | |
|---|---|---|
| d_d_lighting_price_1 | −0.00175008 | 0.7956 |
| d_relative_health_expenditure_5 | 0.167255 | 0.4566 |
| d_urbanization_rate_5 | **0.230642** | **0.0257** |
| d_daily_caloric_intake_2 | 0.00218014 | 0.1905 |
| perc_gdp_4 | 0.00128284 | 0.9424 |
| ww1 | **−1.17242** | **0.0217** |
| ww2 | 0.0980225 | 0.7983 |
| d_d_child_mortality | **−0.431833** | 3.63E-12 |

**Table 9.**

*Model statistics.*

| **Statistic** | **Value** |
|---|---|
| R-squared | 0.512713 |
| Adjusted R-squared | 0.486294 |
| Durbin-Watson | 2.239456 |

Time-series regression analysis results suggest that daily caloric intake, lighting price,relative health expenditure, GDP adjusted for stationarity and World War 2 are not statistically significant with p-values above all thresholds. On the other hand, urbanization rate has a strong influence on life expectancy. If urbanization rate increased 5 years ago, life expectancy would increase by 0.23 years (84 days) today (keeping other independent variables constant). During World War 1 period (1914–1918) life expectancy reduced by 1.17 years (1 year and ~2 months) compared to non-war years, controlling for other factors. When there's an acceleration (second difference) in child mortality of 1 unit, life expectancy decreases by 0.43 years (157 days), holding all other independent variables constant. If there was an acceleration (second difference) in average working hours of 1 unit 3 years ago, life expectancy decreases by 0.005 years ( ~2 days) today, holding all other independent variables constant. About 48.6% (adjusted R-squared result) of the observation variance can be explained with selected variables; this shows that there are opportunities to explore the same relationship with additional variables.

## Conclusions

The objective of this study was to analyzed the relationship between caloric intake and life expectancy in UK over time using time series econometric techniques. As trends and unhealthy diet awareness in United Kingdom rises, it is logical to assume that people fear that gluttony is one of the most universal causes of death. However, our research indicates that this may not be so.

Our time series regression analysis indicates that our main dependant variable, caloric intake, proved to have no statistical significance. Child mortality, on the other hand, had the highest impact, which negatively affected life expectancy. Although some would assume that child mortality is a direct formula of life expectancy, our data established that it is not. There were years when both life expectancy and child mortality increased, indicating that they are not directly correlated. Other variables that impacted life expectancy are WW1, average working hours which have negative effect and urbanization rate which had a positive effect. Out of those, urbanization and working hours we expected to have an influence from the introduction. Interestingly enough, among other control variables, whose results proved to be insignificant included WW2 and relative health expenditure. Other not mentioned variables were insignificant.

The main problems that cause our model to not be the best linear unbiased estimator are normality, heteroscedasticity and autocorelation. In case of normality, we found no fundamental cause, which could be accounted for in our model. Possible bias in p-values can also be attributed to this issue. Due to the lack of available data that would span long enough to increase length of the dataset or to add relevant variables we were also unable to solve the problems of heteroscedasticity and autocorelation. The fact that some part of our data was imputation may also explain biased results. However, there were no significant problems with stationarity, multicolinearity and pseudo-causality, which returns some liability to our model.

Trying to make the dataset more accurate could lead to a better model overall, despite that, improvements would be minor and should not lead to vastly different results. Major improvements in predicting life expectancy could probably only appear from additional variables such as income inequality, air pollution, medicine innovations, child labor participation and others. Finally, higher predicting power of a model could also come from using a different model than Ordinary Least Squares.

## References

Pingali, P. (2012). Green Revolution: Impacts, limits, and the path ahead. Proceedings of the National Academy of Sciences, 109(31), 12302–12308. https://doi.org/10.1073/pnas.0912953109

NHS. (2023, April 17). Understanding Calories. Nhs.uk. https://www.nhs.uk/live-well/healthy-weight/managing-your-weight/understanding-calories/