# A blue and white logo Description automatically generated

# Age Prediction with Multiple Linear Regression

# Executive Summary

Western Govenors University

# Kim Fowler

# 02/03/2025

**Age Prediction with Multiple Linear Regression**

**Executive Summary**

People in the US are living longer. The World Health Organization predicts the world's population over 60 will nearly double from 12% to 22% between 2015 and 2050. (World Health Organization 2024). This increase in population could cause some concern for major life insurance companies. Life insurance provides financial assistance to loved ones (beneficiaries) when the insured person dies. When life expectancy and actual beneficiary survival rates surpass expectations, it leads to cash flow that exceeds projections (Investopedia 2021). This increasing life expectancy phenomenon causing unexpected cash flow is called the longevity risk factor. The longevity risk factor should prompt major life insurance companies to investigate which health and lifestyle factors can predict an individual's age. I propose a research project to address the hypothesis; the variables Height (cm), Weight (kg), Cholesterol Level (mg/dL), BMI, Blood Glucose Level (mg/dL), Stress Levels, Bone Density (g/cm²), Sun Exposure, Gender, Physical Activity Level, Smoking Status, Alcohol Consumption, Diet, and Chronic Diseases statistically significantly affect Age (years).

I gathered the dataset used to test the hypothesis from Kaggle.com. The Human Age Prediction Synthetic Dataset was downloaded as a CSV file and contains 3,000 rows and 26 columns. The dataset is publicly available and contains no restricted information. It includes various health and lifestyle factors for predicting human age and contains categorical and continuous variables. A table of the variables and their types is below.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type** |
| Gender | The persons gender, ‘Male’ or ‘Female’. | Categorical |
| Height (cm) | The height of the individual in centimeters. | Continuous |
| Weight (kg) | The weight of the individual in kilograms. | Continuous |
| Blood Pressure (s/d) | Blood pressure (systolic/diastolic) in mmHg. | Categorical |
| Cholesterol Level (mg/dL) | Cholesterol level in milligrams per deciliter | Continuous |
| BMI | Body Mass Index, calculated from height and weight. | Continuous |
| Blood Glucose Level (mg/dL) | Blood glucose level in milligrams per deciliter. | Continuous |
| Bone Density (g/cm²) | Bone density in grams per square centimeter. | Continuous |
| Vision Sharpness | Vision sharpness on a scale from 0 (blurry) to 100 (perfect). | Continuous |
| Hearing Ability (dB) | Hearing ability in decibels. | Continuous |
| Physical Activity Level | Categorized as 'Low', 'Moderate', or 'High'. | Categorical |
| Smoking Status | Categorical values including 'Never', 'Former', and 'Current'. | Categorical |
| Alcohol Consumption | Frequency of alcohol consumption. | Categorical |
| Diet | Type of diet, categorized as 'Balanced', 'High Protein', 'Low Carb', etc. | Categorical |
| Chronic Disease | Presence of chronic diseases (e.g., diabetes, hypertension). | Categorical |
| Medication Use | Usage of medication. | Categorical |
| Family History | Presence of family history of age-related conditions. | Categorical |
| Cognitive Function | Self-reported cognitive function on a scale from 0 (poor) to 100 (excellent). | Continuous |
| Mental Health Status | Self-reported mental health status on a scale from 0 (poor) to 100 (excellent). | Continuous |
| Sleep Patterns | Average number of sleep hours per night. | Continuous |
| Stress Levels | Self-reported stress levels on a scale from 0 (low) to 100 (high). | Continuous |
| Pollution Exposure | Exposure to pollution measured in arbitrary units. | Continuous |
| Sun Exposure | Average sun exposure in hours per week. | Continuous |
| Education Level | Highest level of education attained. | Categorical |
| Income Level | Annual income in USD. | Continuous |
| Age (years) | The target variable representing the age of the individual. | Continous |

I completed the data analysis process utilizing Python. This versatile and easy-to-use tool provides plenty of libraries for preprocessing the data, generating visualizations, and building a multiple linear regression model. The steps for the data-analysis process are summarized as follows:

* Exploratory Data Analysis – EDA aims to uncover patterns, identify outliers, and gain initial insights into the data. I utilized several *Pandas* functions and *Plotly* visualizations to analyze and investigate the data. Among the methods of investigation used were boxplots to look for outliers, scatter plots to test for linear relationships between the target variable and predictor variables, and displaying the head of the dataset.
* Preprocessing - Data cleaning aims to prepare the data for regression analysis. I cleaned the dataset to guarantee the reliability and accuracy of the analysis findings. I tested the dataset for missing values, duplicate entries, and outliers. Fortunately, no missing values or duplicate entries were found. Outliers were addressed through imputation to maintain the sample size.
* Data Wrangling - Categorical variables were re-expressed before multiple linear regression analysis. A data analyst needs to transform the categorical responses, for example, 'Yes' or 'No', into integers so that calculations and comparisons can be made. The nominal categorical variables Gender, Physical Activity Level, Smoking Status, Alcohol Consumption, Diet, and Chronic Diseases will be re-expressed with one-hot encoding using the *Pandas* *get\_dummies*() function.
* Variance Inflation Factor (VIF) - The dataset must have no multicollinearity; this is the assumption that none of the independent variables are highly correlated with each other. Calculate each independent variable's variance inflation factor (VIF) to test for multicollinearity. The variance inflation factor can be calculated using the `variance\_inflation\_factor` function from the statsmodels library.
* Predictive Modeling - The initial multiple linear regression (MLR) model was developed using the Ordinary Least Squares (OLS) method from the statsmodels library. The OLS method estimates the coefficients of the linear regression model.
* Feature Selection - Feature selection was performed using the wrapper method known as Backward Stepwise Exclusion. In this method, the analyst iteratively removes features with p-values over the accepted value of 0.05, starting with the variable with the highest p-value. Backward stepwise exclusion works well for feature selection to reduce the initial model because the initial model did not contain a large number of explanatory variables.
* Model Evaluation - To evaluate the model, create a Q-Q plot of the residuals. A qqplot provides a good visual assessment of the model's performance. The qqplot plots the distribution of the residuals; if they plot close to a 45-degree line, they are normally distributed.

The initial multiple linear regression (MLR) model was developed using the Ordinary Least Squares (OLS) method from the statsmodels library. The OLS method estimates the coefficients of the linear regression model. The initial model has an Adj. R-Squared value of 0.928. No independent variables were removed because of a high VIF value. Using the backward stepwise exclusion method to iteratively remove variables with a p-value greater than 0.05, variables Gender, Physical Activity Level, Alcohol Consumption, Diet, and Chronic Disease were removed. The team created a reduced model with Cholesterol Level, Blood Glucose Level, Bone Density, Smoking Status, Vision Sharpness, and Hearing Ability. The reduced model has an Adj. R-squared value of 0.928. A q-q plot of residuals was created. The residuals for the reduced model plot close to the 45-degree line. The reduced linear regression model can be used to predict human age with the formula - Age (years) = 73.02 + 0.03 (Cholesterol Level) + 0.04 (Blood Glucose Level) - 25.87 (Bone Density) – 32.61 (Vision Sharpness) + 0.15 (Hearing Ability) - 0.57 (Smoking Status). The OLS results summary utilized to develop the formula is presented below.

A screenshot of a data

Description automatically generated

A screenshot of a computer

Description automatically generated

The limitation of this project is that the data set Human Age Prediction Synthetic Data, while free, saving time and resources in gathering and processing data, was fictitious. The project was theoretical but demonstrated that a model to predict age could be developed using health and lifestyle variables.

With the success of this research project, the next steps would be to gather comprehensive health and lifestyle information on current customers and their beneficiaries. Collect similar health and lifestyle information for new customers and their beneficiaries before creating a new life insurance policy. Then, complete age prediction with multiple linear regression analysis to gain insight into longevity risk.

By using a predictive multiple linear regression model, insights can be gained into whether an individual may be considered a longevity risk. Identifying individuals who are likely to live longer than expected allows life insurance companies to take proactive measures to mitigate unexpected cash flow issues.

**Sources**

World Health Organization. (2024, October 1). *Ageing and Health*. World Health Organization. <https://www.who.int/news-room/fact-sheets/detail/ageing-and-health>

abdullah, M. (2024). *Human Age Prediction Synthetic Dataset*. Kaggle.com. https://www.kaggle.com/datasets/abdullah0a/human-age-prediction-synthetic-dataset

Grammarly. (n.d.). Grammarly. <https://www.grammarly.com/>