The 2nd International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/4/20220548

# Statistical relation between physical inactivity percent and heart disease mortality

**Fanzhi Lin[1,4,†], Junjie Tang[2,†] and Yemin Yang[3,†]**

[1] Department of Statistics, University of Malaya, Kuala Lumpur, Malaysia
[2] High School Attached to East China University of Technology, Shanghai, Chian
[3] Guanghua QiDi education center, Shanghai, China

wid180722@siswa.um.edu.my
[†]These authors contributed equally

**Abstract.** Today, cardiovascular disease is among the leading causes of mortality. On average, someone dies of cardiovascular disease every 36.1 seconds in the United States. According to 2019 statistics, 2,396 people die from cardiovascular disease per day. Analysis of the influencing variables of heart disease mortality can effectively advise people's living habits, prevent the occurrence of heart disease, and lower the risk of disease. This paper analysis the risk factors for heart attack at the county level in the United States. Influencing factors included adult obesity rate, adult smoking rate, diabetes rate, low birthweight rate, excessive drinking rate and physical inactivity rate. Through data pre-processing, descriptive statistical analysis, data visualization and linear regression analysis, a simple linear regression equation between physical inactivity rate and heart disease mortality was finally obtained. It has a slope of 0.7266, an intercept of 0.0767, and the R-squared of linear regression model is 0.539. It is confirmed that physical inactivity rate and heart disease mortality have a strong positive linear relationship.

**Keywords:** physical inactivity percent, heart disease mortality, regression analysis.

## 1. Introduction

Our group wants to analyze these data, we can find that the six factors are relatively high in the incidence of heart disease, so we can consciously remind people with related symptoms to pay attention to and improve in their daily life. The data from our study can be combined to provide more precise preventive measures for specific groups. For example, for smokers, quitting smoking after the age of 50 will reduce the risk of heart disease and prolong life by 5 to 6 years.

Heart disease is a rather frequent circulation ailment that may have a substantial impact on a patient's productivity. With modern medical technology, most heart diseases cannot be cured. Only by taking medicine, hospitalization and other methods to relieve and improve the condition, and often worry about the recurrence of heart disease. Not only that, but all the time recuperation requires a lot of medical bills.

Our group collected an analysis of six factors affecting the incidence of heart disease at the county level in the United States, health pct adult obesity, health pct adult Smoking, health pct diabetes, health pct low birth weight, health pct excessive drinking and health pct physical inactivity. In our survey, these six influencing factors were the main factors leading to the development of heart disease.

DOI: 10.54254/2753-8818/4/20220548

First, obese individuals are at a higher risk for a variety of other major health disorders, including as cardiovascular disease, stroke, diabetes, some malignancies, and poor mental health. Some studies suggest that for some people of Asian descent, obesity-related health risks may occur at a lower body mass index.

Secondly, smoking is harmful to health. In the first place, it affects the respiratory tract and causes chronic respiratory disorders such as bronchitis, chronic bronchitis, chronic obstructive pulmonary disease, and even lung cancer. At the same time, it may also lead to high risk factors for diseases such as heart disease and coronary heart disease. Coronary cardiovascular disease develops when plaque or a clot narrows or obstructs the arteries that provide blood to the heart muscle. The chemicals in cigarette smoke may thicken the blood and cause clots to develop in veins and arteries. A clot obstruction may result in a heart attack and abrupt death.

Thirdly, diabetic heart disease refers to the heart disease complicated by or associated with diabetic patients, including coronary atherosclerotic heart disease (coronary heart disease), diabetic cardiomyopathy, and autonomic nerve disorders. Dysfunction, such as hypertension, can also include hypertensive heart disease.

The birth of a baby weighing less than 5,5 pounds may be an independent risk factor for changes in heart function that can lead to heart failure, a condition in which the heart is unable to pump enough blood and oxygen to meet the body's needs, resulting in a variety of health issues, including heart disease.

Fifthly Excessive drinking has multiple and multifaceted effects on the structure and function of the heart, the electrical conduction system of the heart, and the blood vessels of the heart. Alcohol stimulates and excites the autonomic nerves of the heart. Alcoholic heart disease caused by excessive drinking is mainly caused by the damage to the heart caused by long-term drinking. We know that after drinking alcohol, on the one hand, the heart rate will increase, and the energy and oxygen consumption of myocardial cells will increase, which will cause myocardial strain for a long time, and ethanol can also cause coronary atherosclerosis, myocardial ischemia and hypoxia, and hypertrophy of the heart.

Finally, physical inactivity is not that closely related to heart disease, but according to data studies, physical inactivity is an established risk factor for cardiovascular disease, with about a third of ischemic heart disease associated with physical inactivity.

By predicting heart disease patients, we can well plan medical countermeasures, including the equipment of doctors and equipment, and the storage of related drugs.

## 2. Literature review

Arroyo-Quiroz, Carmen et al. found that the number of patients suffering from coronary heart disease (CHD) in Mexico is continually increasing year after year, with a 48% increase in mortality since 1980. Nevertheless, no research has examined the causes of these increases. They used the previously validated IMPACT model to explain observed variations in CHD mortality among Mexican adults. Data from 2002 and 2012 are used in the model. Results CHD mortality climbed by 33.8 % in males and 22.8 % in women between 2000 and 2012. The IMPACT model explained 71% of the increase in CHD mortality. The increased prevalence of diabetes (43%), physical inactivity (28%), and total cholesterol (24%) in the whole population were the primary causes of the rise in mortality. Together, advances in medical device and surgical care avoided or delayed 40.3% of fatalities; 10% were attributed to advancements in secondary preventive treatment after MI, and 5.3% to community-based heart failure treatment. Arroyo-Quiroz, Carmen, et al. came to the conclusion that that CHD mortality is increasing in Mexico due to growing trends in percentage of diabetes etc and inadequate utilization of CHD treatments. Urgently required are population-level measures to minimize CHD risk factors, as well as enhanced access to and equitable distribution of medications.

The change of systolic blood pressure level has nothing to do with the increase of plasma and cholesterol. Heart disease will increase the incidence rate. Many people ignore the impact of body fat distribution on the incidence rate, which will increase the risk. Among them, there are four risk factors, including central obesity called "syndrome X". There are bad factors in these reasons, which will increase the risk of heart disease. Through the investigation and Research on environmental pollutants

DOI: 10.54254/2753-8818/4/20220548

and heart disease, it is shown that the environment has a great impact on heart disease, and that the environment can change the risk of heart disease when the genes are the same, which can very well prove that the environment is a major factor for heart disease risk. Through a large number of data, it can be found that if parents drink a lot of alcohol, they will increase the probability of heart disease of their children. Although there are differences, on the whole, the amount of alcohol consumed by parents is strongly related to the risk of coronary heart disease in children, which highlights the need to improve health awareness to prevent the incidence rate of heart disease in children.

Timothy et al. discussed the evidence linking environmental pollutants and cardiovascular disease (CVD) at present. A large amount of evidence shows that environmental factors have a great impact on the risk coefficient, incidence rate and severity of cardiovascular diseases. Immigration studies have shown that in genetically stable populations, environmental changes can significantly change the risk of cardiovascular disease. In addition, cardiovascular disease risk is affected by changes in nutrition and lifestyle choices. These findings demonstrate conclusively that persistent environmental stress is a substantial risk factor for cardiovascular disease.

Through research, it has been found that people's personality and social environment are also a very big influencing factor of heart disease. In society, such as being isolated from others, having conflicts with leaders at work, and having conflicts with teachers at school, will have a significant increase in the risk of heart disease. It can be seen that people's personality and social environment are also a very big influencing factor of heart disease, It will have a great impact on the risk of heart disease and death.

The global population has aged significantly in recent years. Tully, Mark A. et al. were concerned that the investigation of the consequences of physical inactivity on health has often neglected older persons. By exploring numerous databases for systematic reviews and/or meta-analyses of longitudinal observational studies, they explored the connection between physical activity and any physical or mental health outcomes in persons aged 60 years. Using AMSTAR, the quality of the contained reviews was evaluated. Physically active older persons (60 years) had a lower risk of all-cause mortality and cardiovascular mortality, Alzheimer's disease, depression, etc. according to the findings[1-10].

Kulhánová, Ivana et al. estimated the effect of smoking, overweight/obesity, and lack of physical activity on IHD mortality among people aged 30-79 years, respectively, at different levels of education, using a method based on population attribution scores, classified by differences in educational attainment. They determined that smoking was the most significant risk factor for males in northern and eastern European communities, whereas obesity and overweight were the most significant risk factors for women in southern European groups. Similarly, physical inactivity had a lesser impact in reducing disparities in IHD mortality compared to smoking and obesity. IHD mortality may be greatly lowered with little education.

Floud Sarah et al. examined the relationship between educational and geographical deprivation and coronary heart disease risk, as well as the influence of smoking, alcohol intake, physical activity, and BMI on these disparities. The first coronary event and CHD mortality in 1,202,983 women with an average age of 56 years were studied. The relative risk of CHD was estimated using Cox regression, which calculated the percent decrease in the adjusted correlation likelihood ratio statistic for each factor separately and in combination to assess the degree of any association that smoking, alcohol, physical inactivity, and BMI could explain. The majority of the relationship between CHD risk and educational and geographical poverty in British women was explained by health-related habits, particularly smoking, and to a lesser extent by alcohol use, physical inactivity, and BMI.

Zhang, Senmao et al. found that if parents drink a lot of alcohol, they will increase the probability of heart disease of their children. Although there are differences, on the whole, the amount of alcohol consumed by parents is strongly related to the risk of CHD in children, emphasizing the need to enhance health consciousness in order to reduce the incidence rate of heart disease in children.

Austin, M.A. et al. investigated the effects of shared genes and shared surroundings on CHD risk variables in 434 pairs of adult female twins in Oakland, California, using genomic and epidemiological analysis. Multiple regression analysis found that the heredity of HDL cholesterol (0.66), LDL cholesterol (0.88), triglycerides (0.53), and relative body weight (0.55) were totally effective, whereas

DOI: 10.54254/2753-8818/4/20220548

the heritability of systolic and diastolic blood pressure (0.42 and 0.25) was not significant. The results of this study help understand why some women had high levels of these CHD risk variables despite adhering to recommended health practices.

By comparing general data from patients with and without CHD, Zhao DH et al. discovered that the proportion of CHD patients with concomitant MB was considerably greater than that of the control group. Multivariate logistic regression analysis was performed showing that MB thickness, systolic and diastolic compression, and MCA systolic stenosis are independent determining variables for MB-related CHD [1-10].

## 3. Dataset

This dataset comes from Kaggle. This dataset contains 3,918 samples, and each sample has seven attributes. These features are adult obesity rate, adult smoking rate, diabetes rate, low birthweight rate, excessive drinking rate, physical inactivity rate, and heart disease mortality. There are some missing values in the data. And after removing rows with missing data, the total number of samples is 2109. Specifically, the mean of physical inactivity is 0.268 and the standard deviation is 0.0548, indicating that the data are not particularly discrete. Maximum value is 0.442, minimum value is 0.09. The quartiles are 0.231, 0.268, 0.306. The mean heart disease mortality is 0.2712 and the standard deviation is 0.0542. Maximum is 0.512 while minimum is 0.135. The quartiles are 0.231, 0.264, and 0.306.

## 4. Implementation

First, read and understand the data. Use the pandas library to import data and understand the structure of the data. Next, visualize the data. Data visualization is curial to discover the data. And scatter plots can effectively illustrate the link between two variables. Therefore, define the heart disease mortality as the dependent variable and the other features as the independent variables. Use seaborn to do the visualization. The scatter plot is shown in figure 1. It reveals that heart disease mortality is negatively correlated with excessive drinking percentage and positively correlated with the other features.

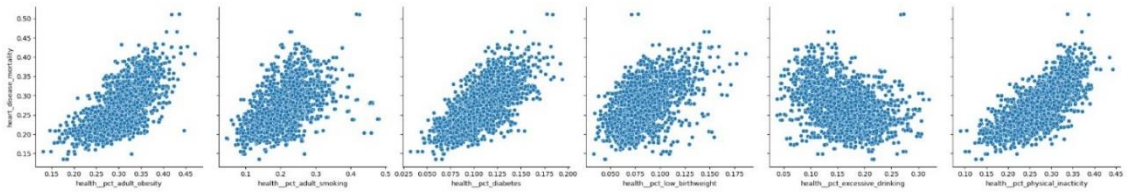


**Figure 1.** Scatter graph.

Although the scatter plot can be used to observe the distribution, it cannot be used to get an exact correlation coefficient. Consequently, using heatmap to get and display the correlation coefficient between each feature. The strongest association coefficient, 0.73, was observed between heart disease mortality and physical inactivity percentage.

DOI: 10.54254/2753-8818/4/20220548

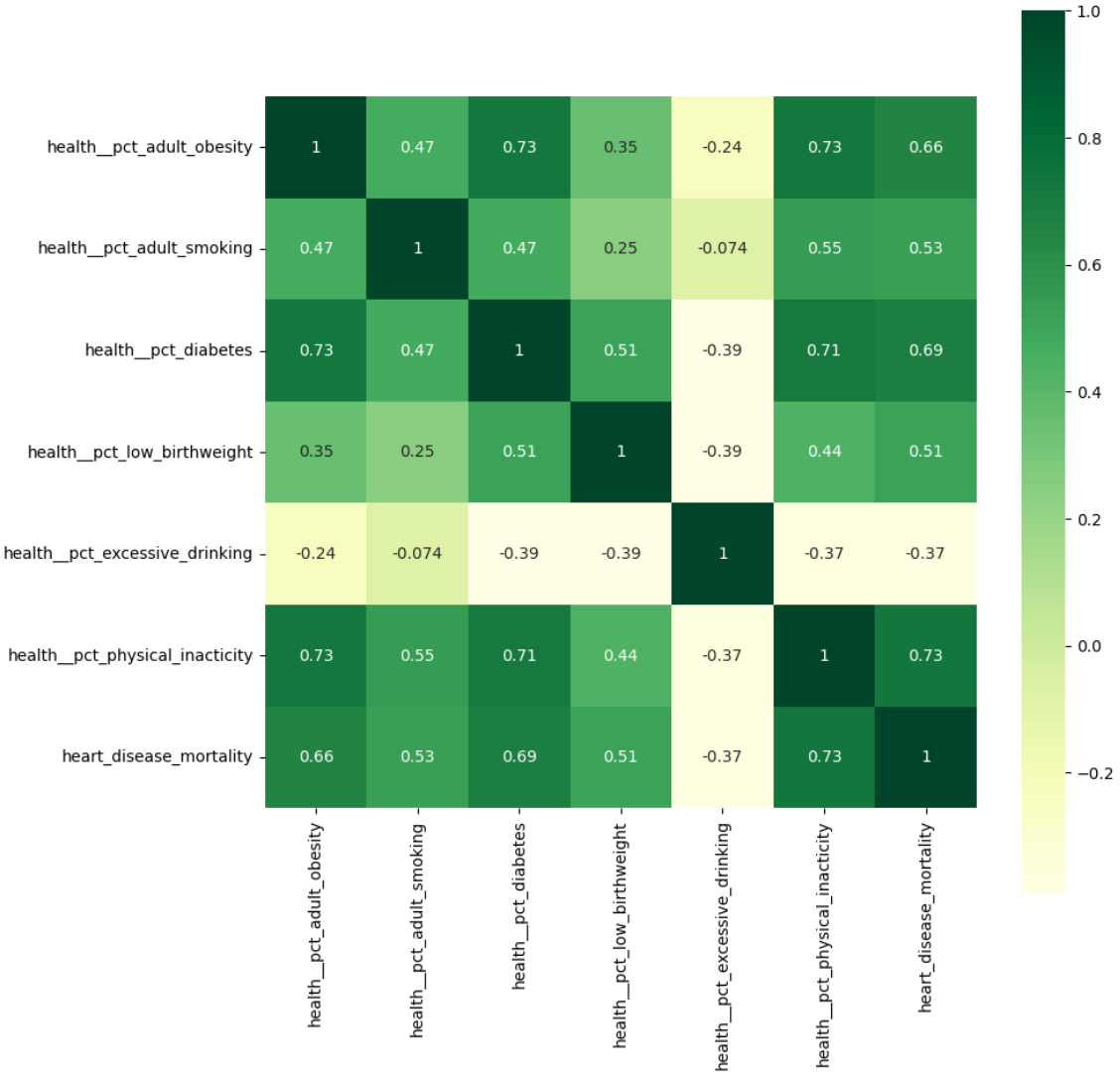


**Figure 2.** Heatmap.

After that, perform a simple linear regression. Assign feature variables and split variables to training and test sets. It is usually a good practice to keep 70% of the data in the training dataset and the remaining 30% in the test dataset. Build a linear model. Import the statsmodel.api library that performs linear regression. By default, the statsmodels library fits a line through the origin on the dataset. But in order to have an intercept, the add_constant property of statsmodels needs to be used manually. After adding the constants to the X_train dataset, proceed to fit a regression line using the OLS (Ordinary Least Squares) property of statsmodels.

## 5. Methodology

The objective of regression analysis is to examine the quantitative relationship between variables and describe this relationship using a specific mathematical expression in order to assess the impact to which changes in one or more variables (independent variables) have an impact on another specific variable (dependent variable). To be more detailed, regression analysis is specifically useful for addressing the following issues: Figure out the mathematical relationship between the variables using the data from the sample set as a starting point; Perform various statistical tests on the veracity of these correlations and determine which of the numerous variables that influence the dependent variable are significant and

DOI: 10.54254/2753-8818/4/20220548

which are insignificant; With the help of the relational expression that was found, the value of the dependent variable can be estimated or forecasted based on the value of one or more independent variables, and then the accuracy of the estimation or prediction can be discussed.

The simplest model in regression analysis is a univariate linear regression model, commonly referred to as a simple linear regression model, which has just one dependent variable and one independent variable.

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{1}$$

The population regression function is the name given to the formula above. The formula's unknown parameters, also known as regression coefficients, are denoted by $\beta_1$ and $\beta_2$. The $t_{th}$ observation of Y and X, respectively, are denoted by $Y_t$ and $X_t$. $u_t$ is a specific random variable that represents the influence of several other factors not included in the equation on Y. It is sometimes referred to as a random interference term or a random error term.

In reality, it is impossible to comprehend all the values of the variable population since the overall unit number of the phenomena is typically large, and in many cases even infinite. It must be estimated using the sample's information. The line of a simple linear regression model can be expressed as:

$$\widehat{Y_t} = \widehat{\beta_1} + \widehat{\beta_2} X_t \tag{2}$$

In the formula, $\widehat{Y_t}$ represents the Y value corresponding to $X_t$ on the sample regression line, which can be considered an estimate of $E(Y_t)$. $\widehat{\beta_1}$ is the intercept coefficient, $\widehat{\beta_2}$ is the slope coefficient, and they are estimates of the overall regression coefficients $\beta_1$ and $\beta_2$. The actual observed value of the dependent variable $Y_t$ differs from $\widehat{Y_t}$. If you use $e_t$ to indicate the difference between the two, $e_t = Y_t - \widehat{Y_t}$, there are the following formula:

$$Y_t = \widehat{\beta_1} + \widehat{\beta_2} X_t + e_t \ \ (t=1,2,3,\ldots,n) \tag{3}$$

The above equation is known as the sample regression function, where $e_t$ is called residual. The sample regression function represents the population regression function approximatively. The primary objective of regression analysis is to fully utilize the information provided by the sample such that the sample regression function is as close as feasible to the actual overall regression function.

Least squares is a method for estimating regression coefficients in which the sum of squared residuals is minimized. Assume:

$$Q = \sum e_t^2 = \sum \left(Y_t - \widehat{Y_t}\right)^2 = \sum \left(Y_t - \widehat{\beta_1} - \widehat{\beta_2} X_t\right)^2 \tag{4}$$

Take to zero the partial derivative of Q. Following sorting and resolution, we have

$$\widehat{\beta_2} = \frac{n\sum X_t Y_t - \sum X_t \sum Y_t}{n\sum X_t^2 - (\sum X_t)^2} \tag{5}$$

$$\widehat{\beta_1} = \overline{Y} - \widehat{\beta_2}\overline{X} \tag{6}$$

The mathematical formula for estimating the regression coefficient of a population using the least-squares approach is a function of the sample observations and is often referred to as the least-squares estimator.

The least squares estimator is a linear function of $Y_t$ whose expected value equals the actual value of the regression coefficients for the population. Therefore, the least squares estimator is a linear, unbiased estimator of the regression coefficients for the population. It is also possible to demonstrate that, among all linear unbiased estimators, the least squares estimator of regression coefficients has the minimum variance; as the sample size grows, this variance will continue to decrease. In other words, the least squares estimator of regression coefficients is both the ideal linear unbiased estimator and a consistent estimator.

DOI: 10.54254/2753-8818/4/20220548

According to the principle of parameter interval estimation, the following regression coefficient interval estimation formula can be obtained

$$\widehat{\beta_j} \pm t_{\frac{a}{2}}^{(n-2)} \times S_{\widehat{\beta_j}} \quad (j=1,2) \tag{7}$$

where $S_{\widehat{\beta_j}}$ is the sample standard error of the regression coefficient estimation, $t_{\frac{a}{2}}^{(n-2)}$ is the two-sided critical value of the $t$ distribution with the significance level $\alpha$ and the degrees of freedom $n-2$.

$$S_{\widehat{\beta_1}} = S\sqrt{\frac{1}{n} + \frac{\bar{X}}{\sum(X_t-\bar{X})^2}} \tag{8}$$

$$S_{\widehat{\beta_2}} = \frac{S}{\sqrt{\sum(X_t-\bar{X})^2}} \tag{9}$$

The determination coefficient ($r^2$) is an all-encompassing measure of the regression model's fit. The higher the degree of model fit, the larger the determination coefficient.

$$r^2 = \frac{\sum(\widehat{Y_t}-\bar{Y})^2}{\sum(Y_t-\bar{Y})^2} = 1 - \frac{\sum(Y_t-\widehat{Y_t})^2}{\sum(Y_t-\bar{Y})^2} \tag{10}$$

In regression analysis, the significance test comprises two components: the significance test of each regression coefficient and the significance test of the overall regression equation. Since there is only one explanatory variable in a simple linear regression model, the test for $\beta_2=0$ is similar to the significance test for the full equation. The significance test of the regression coefficient is to test the relevant hypothesis of the overall regression coefficient according to the results of the sample estimation. The tests for $\beta 1$ and $\beta 2$ are the same, but the test for $\beta 2$ is more important because it shows how much the independent variable affects the dependent variable.

T test:

$$H_0: \beta_2 = \beta_2^*, H_1: \beta_2 \neq \beta_2^* \tag{11}$$

where $H_0$ is the null hypothesis, $H_1$ is the alternative hypothesis, and $\beta_2^*$ is the actual value of the assumed population regression coefficient. $\beta_2^* = 0$ is frequently set in various computer algorithms for regression analysis. Because the value of $\beta_2^*$ can show if X has a major effect on Y.

Determine the significance level $\alpha$. The size of the significance level should be determined according to the size of the loss that may result from which type of error is made. In general, 0.05 is acceptable.

Calculate t-values for regression coefficients

$$t_{\widehat{\beta_2}} = \frac{\widehat{\beta_2}-\beta_2^*}{S_{\widehat{\beta_2}}} \tag{12}$$

Check the t distribution table using $\alpha$ and df to discover the crucial value. Accept the alternative hypothesis and reject the null hypothesis if the absolute value of $t_{\widehat{\beta_2}}$ is larger than the absolute value of the crucial value; else, accept the null hypothesis.

The p test may also be used to evaluate the significance of the regression coefficient. The first three steps are identical to the t test; however, after the t value is calculated, it is not compared to the critical value of the t distribution; rather, the probability that the t statistic with degree of freedom $n-2$ is greater or less than the $t_{\widehat{\beta_2}}$ calculated from the sample observations, also called p value. Then compare it with the given significance level $\alpha$, if $p$ is less than $\alpha$, reject the null hypothesis, otherwise accept the null hypothesis.

Result

DOI: 10.54254/2753-8818/4/20220548

A simple linear regression analysis using OLS model with the physical inactivity percentage as the independent variable and heart disease mortality as the dependent variable is performed. In the final result, $R-squared$ is 0.539 and means 53.9% variation in heart disease is explained by physical inactivity percentage. R-square demonstrates that our regression is feasible. The $Adj.\ R-squared$ is 0.538. The constant term after our regression analysis is 0.0767, and the slope is 0.7266. Since the p-value is 0.000 which is less than 0.05 and there is no 0 inside the confidence interval which is [0.025, 0.975], the null hypothesis is rejected in favour of the alternative hypothesis. The figure 3 below shows the regression line.

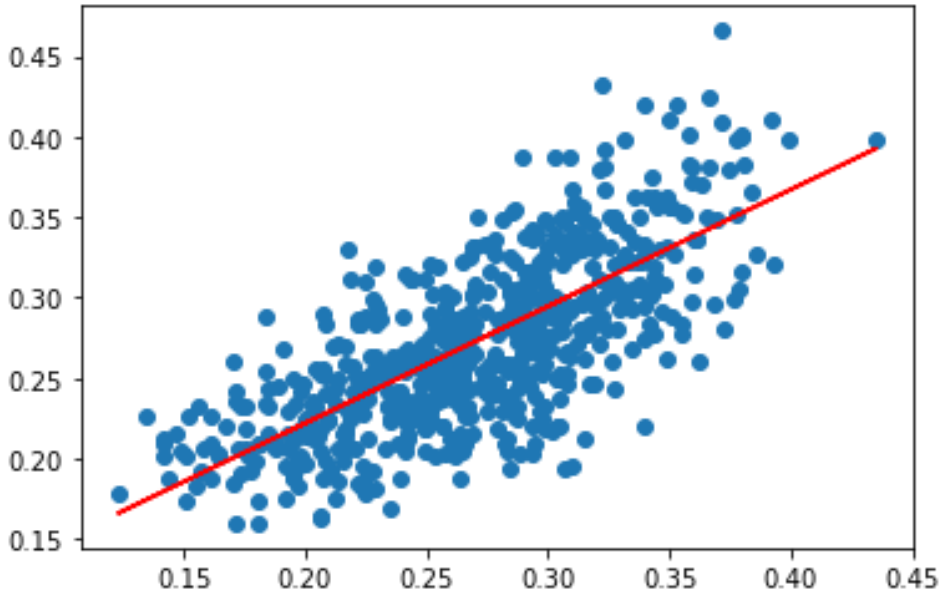


**Figure 3.** Linear regression line.

## 6. Conclusion

Influencing factors of heart disease at county level in the United States were analysed. The original 3198 groups were recorded through data pre-processing. However, due to some missing data, there are still 2109 groups of data left through data comparison and deletion. After integrating all the data into jupyter notebook (anaconda3), descriptive statistical analysis was carried out on these data, and through data visualization, six parts of adult obesity percentage data (adult smoking percentage, diabetes percentage, low weight percentage, excessive drinking percentage and lack of physical exercise percentage) were clearly expressed in the form of images, Then, linear regression and other steps are used to establish a linear model for the dependent variable (heart disease mortality) and the dependent variable (lack of physical exercise percentage). From the above chart analysis, it can be concluded that obesity, smoking, excessive drinking, poor physical fitness and other factors will increase the incidence rate of heart disease. At the same time, studies have shown that when plasma cholesterol levels increase, the risk of heart disease will also increase synchronously. A large amount of evidence shows that the environment has a great impact on the risk coefficient, incidence rate and severity of cardiovascular diseases. In addition, cardiovascular disease risk is affected by changes in nutrition and lifestyle choices. These data can well and effectively prove the view that chronic environmental stress is an important determinant of cardiovascular disease risk. In the survey of the parents of heart disease patients, it was found that most of their parents drank a lot of alcohol, which showed that the amount of alcohol consumption of parents was also significantly related to the prevalence of heart disease. Therefore, we hope that our family can strengthen exercise, live a healthy and green life, reduce the incidence rate of heart disease, and avoid the devastation of heart disease.

## References

[1] Arroyo-Quiroz, Carmen, et al. "Explaining the Increment in Coronary Heart Disease Mortality in Mexico between 2000 and 2012." PLoS ONE 15.12 (2020): 1-15. Print.

[2] Austin, M. A., et al. "Risk Factors for Coronary Heart Disease in Adult Female Twins: Genetic Heritability and Shared Environmental Influences." American Journal of Epidemiology 125.2 (1987): 308-18-18. Print.

[3] Cunningham, Conor, et al. "Consequences of Physical Inactivity in Older Adults: A Systematic Review of Reviews and Meta‐Analyses." Scandinavian Journal of Medicine & Science in Sports 30.5 (2020): 816-27. Print.

[4] Floud, Sarah, et al. "The Role of Health-Related Behavioural Factors in Accounting for Inequalities in Coronary Heart Disease Risk by Education and Area Deprivation: Prospective Study of 1.2 Million Uk Women." BMC Medicine 14 (2016): 1-9. Print.

[5] Kulhánová, Ivana, et al. "The Role of Three Lifestyle Risk Factors in Reducing Educational Differences in Ischaemic Heart Disease Mortality in Europe." European Journal of Public Health 27.2 (2017): 203-10. Print.

[6] Smith, T. W., and J. M. Ruiz. "Psychosocial Influences on the Development and Course of Coronary Heart Disease: Current Status and Implications for Research and Practice." 2002: 548-68. Print.

[7] Tie-Ning, Zhang, et al. "Environmental Risk Factors and Congenital Heart Disease: An Umbrella Review of 165 Systematic Reviews and Meta-Analyses with More Than 120 Million Participants." Frontiers in Cardiovascular Medicine 8 (2021). Print.

[8] Wilcox, I., et al. "'Syndrome Z': The Interaction of Sleep Apnoea, Vascular Risk Factors and Heart Disease." 1998: S25-S28. Print.

[9] Wong, S. F., et al. "Factors Influencing the Prenatal Detection of Structural Congenital Heart Diseases." Ultrasound in Obstetrics and Gynecology 21.1 (2003): 19-25-25. Print.

[10] Zhang, Senmao, et al. "Parental Alcohol Consumption and the Risk of Congenital Heart Diseases in Offspring: An Updated Systematic Review and Meta-Analysis." EUROPEAN JOURNAL OF PREVENTIVE CARDIOLOGY 27.4 (2020): 410-21. Print.

[11] Zhao, Dong-Hui, et al. "Myocardial Bridge-Related Coronary Heart Disease: Independent Influencing Factors and Their Predicting Value." World journal of clinical cases 7.15 (2019): 1986-95. Print.