Determinants of Heart Disease: A Multi-Site International Study

By: Zhuowen Dai, Cong Liu, Fengyuan Tang, Guangye Tao, Lingrui Yu

ECO372: Data Analysis and Applied Econometrics in Practice

University of Toronto

Department of Economics

Group Project

December 2021

1. Introduction

The high prevalence and incidence of heart disease have become a deep concern for the public health sector. According to a 2017 report published by the Public Health Agency of Canada, heart disease is the second leading cause of death after cancer, and a leading cause of hospitalization in the country (Public Health Agency of Canada, 2017). The damages from heart syndromes greatly impair human life and increase their risk of mortality. The cost of heart disease is also considerable in economic terms. Specifically, chronic heart failure consumes 1–2% of the total health care budget in developed countries (Berry et al., 2001). It would be insightful for public health to identify key risk factors behind heart diseases so that early detection of heart disease patients is possible which is beneficial to social welfare.

2. The Context and Data

2.1. Literature Review

Some theories have been established regarding the risk factors for heart disease. One of the frequently seen and well-known findings is the association between hypertension and coronary heart disease (CHD). Hypertension is clinically defined as the condition of systolic blood pressure higher than 130 mm Hg or diastolic blood pressure higher than 80 mm Hg for adults (Byrd & Brook, 2019). The PROCAM study, which aims to determine coronary heart disease factors, tracked a set of participants and analyzed the panel data collected over time. Follow-up on the male study participants aged 40 to 65 years who were free of heart conditions shows that hypertension and hyperlipidemia are independent risk factors for CHD (Assmann & Schulte, 1988). In addition, high serum cholesterol level is associated with a poorer prognosis in coronary artery disease. A study on 133 patients hospitalized due to progressive heart failure found that higher cholesterol level indicates worse outcomes for patients with coronary artery disease (Sakatani et al., 2005). Previous findings also showed that the age-adjusted prevalence of myocardial ischemia and myocardial infarction was higher for individuals with hypertension, hypercholesterolemia and hyperglycemia than in normal subjects (Kim et al., 1996). Apart from observational studies, a previous experimental study using a randomized placebo-controlled trial confirmed that a high heart rate is a risk factor for heart failure (Böhm et al., 2010).

However, most past studies were based on early-year patient data. The time gap is evident between the research and the current heart disease data. The findings of past studies are likely no longer applicable today. Therefore, it is worth studying the current disease dataset to observe whether the risk factors are associated with heart disease and whether their contribution has changed over time.

2.2. Data

Hospital records sourced from the United States (Cleveland and Long Beach), Switzerland and Hungary were available as independent datasets in the UCI Machine Learning Repository. Later consolidated, merged and posted on Kaggle.com, we accessed this multi-regional "Heart Failure Prediction" dataset. Referring to what past researches have studied, we decided to focus our analysis on the following factors available in the data to investigate risk factors associated with heart disease: individual's age, sex, experience of exercise-induced angina, resting blood pressure, serum cholesterol, fasting blood sugar, and maximum heart rate achieved.

First, we generated a heatplot to illustrate the correlation between the factor variables. From Plot 1, heart disease is positively correlated with age, male gender, resting blood pressure, and fasting blood sugar. However, heart disease has a negative correlation with serum cholesterol and maximum heart rate. The correlation between maximum heart rate and heart disease is -0.4, which is relatively large compared to other correlations.

Based on the numeric summary of data (Table 1), men comprise approximately 79% of the sample. The average age of the sample is about 53.5 years old. Moreover, around 40.41% of patients have experienced angina induced by exercise before, and more than half of the patients are diagnosed with heart disease. To better examine how hypertension influences heart disease risk, we generated a dummy variable based on resting blood pressure indicating whether an individual has hypertension. The blood pressure threshold is set at 130 mmHg according to the clinical definition of hypertension (Byrd & Brook, 2019). Table 1 shows that about 59% of the individuals had high blood pressure issues. In addition, cholesterol has many zero values, which may explain its negative correlation with heart disease in the heatplot. To address this issue, we replaced all zero non-reporting cholesterol observations with the sample mean cholesterol level. The mean imputation gave us a sensible set of cholesterol measures.

We then looked at the mean differences in our factors of interest between individuals with and without heart disease. From Table 2, there are 508 individuals diagnosed with heart diseases. The mean age of heart disease patients is about five years older than the "healthy" individuals. Heart disease patients also have 25% more men. We also noticed that those diagnosed with heart disease have a 0.23 mg/dl higher fasting blood sugar. Heart disease patients also achieve, on average, 20.5 beats per minute higher maximum heart rate. The mean differences are consistent with the correlation heatplot.

3. Regression analysis

3.1. Multiple Linear Regression

The first model for the relationship between heart disease and selected factors is a linear probability model. The specification takes the following form:

$$HeartDisease_i = \beta \times RegVar_i + \epsilon_i \, ,$$

where $RegVar_i$ are factors listed above: individual age, sex, experience of exercise-induced angina, resting blood pressure, serum cholesterol, fasting blood sugar, and maximum heart rate achieved. Most regression variables are estimated to have a significant effect on heart disease probability, with the exception of hypertension and serum cholesterol. There is a positive association between age and the probability of getting heart disease, which means the older the age, the higher the probability of getting heart disease. Under the same condition, males are 21.8 percentage points more likely to have heart disease than females. Moreover, there is a positive association between fasting blood sugar and the probability of getting heart disease. Individuals with an elevated level of fasting blood sugar are estimated to have a higher probability of getting heart disease by 21.7 percentage points. By contrast, there is a negative association between the maximum heart rate and the probability of getting heart disease. The higher the maximum heart rate, the lower the probability of getting heart disease.

3.2. Heterogeneity in Heart Disease Risk

The above linear probability model has revealed that fasting blood sugar level has a significant positive association with heart disease occurrence. Interestingly, the effect of hypertension and serum cholesterol on heart disease probability is not statistically significant. Furthermore, the model estimated that hypertension negatively affects the probability of being diagnosed with heart diseases, which contrasts with the earlier literature. Meanwhile, the significant positive effect associated with exercise-induced angina on heart disease probability captured our attention.

The analysis now turns to exploring how these estimated effects vary by individual sex, guided by a common interest in how heart disease risk factors affect men and women differently. A difference-in-differences model is then employed to investigate the sex heterogeneity in the effect of heart disease risk factors. The model focuses on examining the differential effects between sex and hypertension, blood sugar, and exercise-induced angina respectively on heart disease probability, thus incorporating these respective interaction terms.

The interaction between sex and hypertension is the only statistically significant one (t-statistics ≈-2.58) amongst the three interaction terms. Holding all other factors constant, male individuals with hypertension are estimated to be about 15.9 percentage points less likely to develop heart disease compared to women. Likewise, men who have an elevated blood sugar level seem to face a lower risk of heart disease than women. The average reduction in heart disease probability is estimated to be around 9.21 percentage points. In contrast, for individuals who experienced exercise-induced angina, men are marginally more likely to be diagnosed with heart disease, with an estimated increase of 0.29 percentage points.

The addition of interaction terms in the model specification also leads to a surprising change in the sign of the coefficient estimate on hypertension. The interaction regression displays that individuals with high blood pressure are, on average, 9.23 percentage points more likely to have heart disease, holding all other factors constant. Estimation of the effect of hypertension on the heart disease risk from the extended model is consistent with previous findings in the literature (Assmann & Schulte, 1988).

4. Discussion of results

A preliminary analysis of the relationship between potential risk factors and heart disease employs a multiple linear regression on selected physiological measures. MLR results show that men face a 21.8 percentage point higher risk of heart disease. Fasting blood sugar level is positively correlated with heart disease occurrence with statistical significance. Individuals with an elevated blood sugar level are estimated to be 21.7 percentage points more likely to be diagnosed with heart disease. Surprisingly, the initial model estimates a negative relationship between hypertension and heart disease which contradicts with earlier papers.

Due to shared research interests, the group moves on to investigate the potential heterogeneity of heart disease risk factors with respect to individual sex. The extended model estimates a significant interaction effect between sex and hypertension on the heart disease probability. All factors being fixed, men with hypertension are estimated to be about 15.9 percentage points less likely to develop heart disease compared to women. This result is insightful for the public health sector when proposing preventive measures. The sex heterogeneity implies that females with hypertension need a closer track on health conditions to mitigate heart disease risk. Other noteworthy independent factors include age, high fasting blood sugar and experience of angina due to exercise since the regression estimates a significantly positive association between them and heart disease probability.

Relating to the literature findings, the hypertension coefficient estimate becomes positive in this model. Specifically, when holding all other factors constant, the regression showed that hypertensive individuals are, on average, 9.23 percentage points more likely to have heart disease. The new estimation of the effect of hypertension on heart disease is in alignment with earlier papers. A previous longitudinal study concludes that hypertension is one of the independent risk factors of coronary heart disease (Assmann & Schulte, 1988). Past research also suggested that higher cholesterol levels lead to worse outcomes for patients who already have heart disease (Sakatani et al., 2005). Although its estimated effect is marginal, the regression does show a positive association between serum cholesterol and heart disease probability.

5. Limitation of results

The dataset provides limited information on individual demographic features. Common covariates include but are not limited to ethnicity, education level and income level, yet none of them are available in the data. The analysis cannot control for their influence on heart disease risk. There are other health factors that truly underpin the individual risk level of developing heart disease yet are also omitted in the above model specifications. For example, past research suggested that autoimmune diseases may induce a higher risk of heart failure (Kim et al., 2017). It is more commonly known that individual lifestyle and dietary choices greatly affect their risk of heart disease. There is evidence from a cross-cultural cohort study that a healthy diet and lifestyle are essential to the prevention of coronary heart disease (Kromhout et al., 2002). Culture greatly affects individuals: Americans tend to eat and live differently from Hungarians, while people within the same culture tend to share similar lifestyles and diets. However, model specifications in this analysis failed to take into account all individual and region-specific differences. The above models also do not ensure that all fitted values fall within the range for the probability from zero to one.

Possibly due to non-reporting, the data has numerous zero cholesterol records. Although mean imputation is a simple solution to this issue, imputing values at the distribution centre artificially reduces variation in cholesterol measures (UCLA: Statistical Consulting Group, n.d.). As a result, the coefficient estimate on serum cholesterol is likely to be biased. Another data insufficiency is that it only collected the maximum heart rate ever achieved, the model shows inconsistency with the past literature on how resting heart rate affects the risk of heart disease (Böhm et al., 2010): the regression estimates a negative association between maximum heart rate and heart disease probability. Therefore, it is not appropriate to draw a conclusion on the impact of heart rate on heart disease outcome with our research.

6. Conclusion

The analysis set out to explore potential determinants of heart disease and how they affect the disease risk level. Via linear regression models, we obtain evidence from the multi-site data that men and older individuals are more likely to develop heart disease. Hypertension, elevated blood sugar level and previous experience of exercise-induced angina also contribute to heart disease. Most importantly, regression reveals that at the same level of blood pressure, females are exposed to a higher risk of heart disease compared to males.

Our results may provide some insight into public health in terms of reducing the severity of heart disease. We recommend that governments encourage, or even incentivize citizens with monetary rewards to have physical examinations on a regular basis so that blood pressure and blood sugar levels are monitored. Tracking potential abnormalities in these biomarkers enables the detection of heart disease at an earlier stage. It is also recommended for the government to provide health subsidies to women with high blood pressure and men with exercise induced-angina because they are more likely to develop heart disease according to our results. It is usually costly for most people to treat heart diseases. The government may consider offering more funding to scientists to invent new medication on abnormal serum cholesterol, blood pressure and blood sugar level so that disease treatment could be more accessible. More research is also needed to further study the relationship between hypertension, high blood sugar and heart disease for a better understanding of the disease.

References

Assmann, G., & Schulte, H. (1988). The prospective cardiovascular münster (PROCAM) study: Prevalence of hyperlipidemia in persons with hypertension and/or diabetes mellitus and the relationship to coronary heart disease. *American Heart Journal*, *116*(6), 1713–1724. https://doi.org/10.1016/0002-8703(88)90220-7

Berry, C., Murdoch, D. R., & McMurray, J. J. V. (2001). Economics of chronic heart failure. *European Journal of Heart Failure*, *3*(3), 283–291. https://doi.org/10.1016/s1388-9842(01)00123-4

Böhm, M., Swedberg, K., Komajda, M., Borer, J. S., Ford, I., Dubost-Brama, A., Lerebours, G., & Tavazzi, L. (2010). Heart rate as a risk factor in chronic heart failure (SHIFT): The association between heart rate and outcomes in a randomised placebo-controlled trial. *The Lancet*, *376*(9744), 886–894. https://doi.org/10.1016/s0140-6736(10)61259-7

Byrd, J. B., & Brook, R. D. (2019). Hypertension. *Annals of Internal Medicine*, *170*(9). https://doi.org/10.7326/aitc201905070

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Retrieved from: http://archive.ics.uci.edu/ml

fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Kim, C. H., Tofovic, D., Chami, T., Al-Kindi, S. G., & Oliveira, G. H. (2017). Subtypes of heart failure in autoimmune diseases. *Journal of Cardiac Failure*, *23*(8).

https://doi.org/10.1016/j.cardfail.2017.07.044

Kim, S. K., Roh, S. C., Son, J. I., & Choi, B. Y. (1996). Analysis on the relationships among the total cholesterol, fasting blood sugar, hypertension and ischemic heart disease on EKG findings. *Journal of Preventive Medicine and Public Health*, *29*(4), 705-719.

Kromhout, D., Menotti, A., Kesteloot, H., & Sans, S. (2002). Prevention of coronary heart disease by diet and lifestyle. *Circulation*, *105*(7), 893–898. https://doi.org/10.1161/hc0702.103728

Public Health Agency of Canada. (2017, July 18). *Heart disease in Canada: Highlights from the Canadian Chronic Disease Surveillance System*. Government of Canada. Retrieved from https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada-fact-sheet.html.

UCLA: Statistical Consulting Group. (n.d.). *MULTIPLE IMPUTATION IN STATA*. Retrieved from https://stats.oarc.ucla.edu/stata/seminars/mi_in_stata_pt1_new/

Table 1: Summary Statistics of variables in the regression analysis

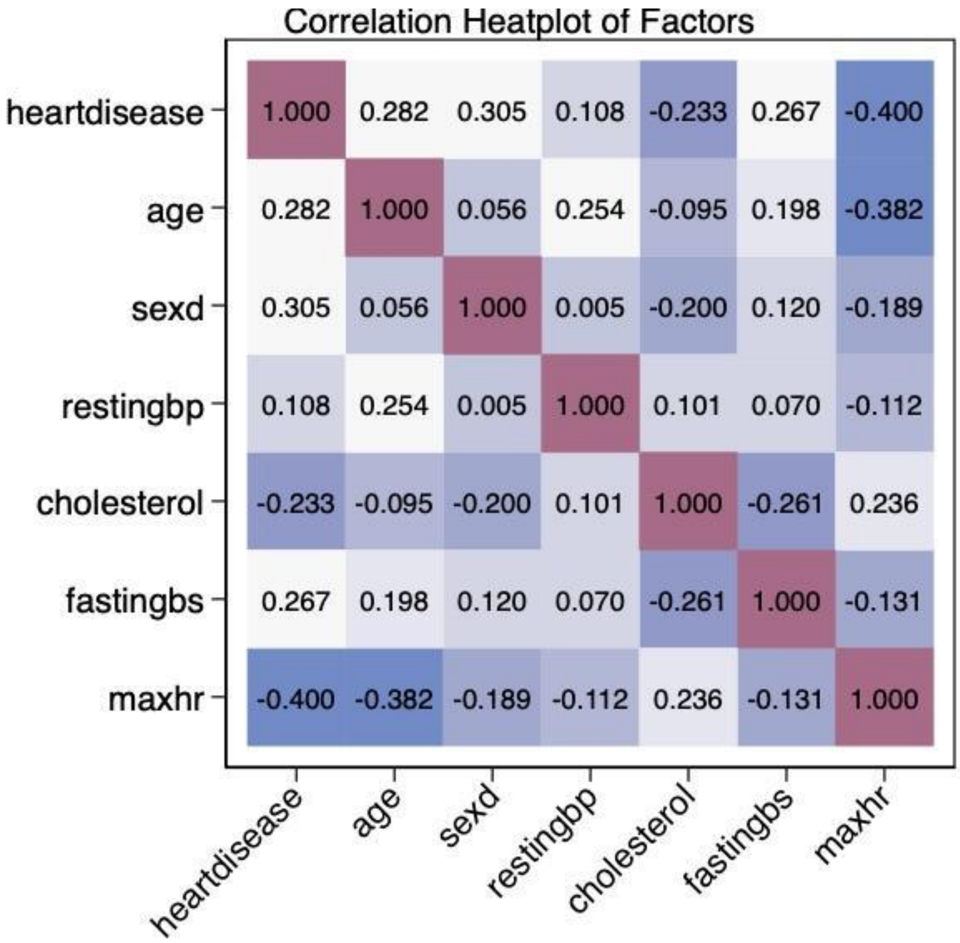
	count	mean	sd	min	max
Age	918	53.51089	9.432617	28	77
Sex (Male = 1)	918	.7897603	.4077009	0	1
Hypertension	918	.5915033	.4918238	0	1
Cholesterol	918	236.0474	56.24095	85	603
FastingGS	918	.2331155	.4230456	0	1
MaxHR	918	136.8094	25.46033	60	202
Exercise-induced Angina	918	.4041394	.4909922	0	1

Table 2: Summary Statistics of variables in the regression analysis by heart disease outcome

	(1)		(2)		(3)	
	Heart Disease		No Heart Disease		Difference	
	mean	sd	mean	sd	b	t
HeartDisease	1.00	0.00	0.00	0.00	-1.00	(.)
Age	55.90	8.73	50.55	9.44	-5.35***	(-8.82)
Sex (Male = 1)	0.90	0.30	0.65	0.48	-0.25***	(-9.26)
Hypertension	0.62	0.49	0.56	0.50	-0.06	(-1.82)
Cholesterol	235.42	57.50	236.82	54.71	1.40	(0.38)
FastingBS	0.33	0.47	0.11	0.31	-0.23***	(-8.76)
MaxHR	127.66	23.39	148.15	23.29	20.50***	(13.23)
Exercise-Induced Angina	0.62	0.49	0.13	0.34	-0.49***	(-17.84)
Observations	508		410		918	

Note: ***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Plot 1: Correlation heat plot between heart disease rate and selected variables



	heartdisease	age	sexd	restingbp	cholesterol	fastingbs	maxhr
heartdisease	1.000	0.282	0.305	0.108	-0.233	0.267	-0.400
age	0.282	1.000	0.056	0.254	-0.095	0.198	-0.382
sexd	0.305	0.056	1.000	0.005	-0.200	0.120	-0.189
restingbp	0.108	0.254	0.005	1.000	0.101	0.070	-0.112
cholesterol	-0.233	-0.095	-0.200	0.101	1.000	-0.261	0.236
fastingbs	0.267	0.198	0.120	0.070	-0.261	1.000	-0.131
maxhr	-0.400	-0.382	-0.189	-0.112	0.236	-0.131	1.000

Table 3: Individual Physiological Measures and Heart Disease Probability

	(1)	(2)
	HeartDisease	HeartDisease
Age	0.00513**	0.00504**
	(0.00156)	(0.00155)
Sex (Male = 1)	0.218***	0.324***
	(0.0338)	(0.0447)
Hypertension	-0.0337	0.0923
	(0.0275)	(0.0531)
Cholesterol	0.000208	0.000175
	(0.000243)	(0.000240)
FastingBS	0.217***	0.297**
	(0.0317)	(0.103)
MaxHR	-0.00339***	-0.00341***
	(0.000620)	(0.000618)
Exercise-Induced Angina	0.372***	0.368***
	(0.0305)	(0.0812)
Sex_Hypertension		-0.159*
		(0.0615)
Sex_FastingBS		-0.0921
		(0.107)
Sex_Exercise-induced angina		0.00288
		(0.0860)
Constant	0.340*	0.272
	(0.149)	(0.148)
Interaction	No	Yes
N	918	918
r2	0.380	0.385

The dependent variable in all regressions is an indicator denoting whether the individual is diagnosed with heart disease. Independent variables include main biomarkers. Robust standard errors reported in parentheses for all regressions.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$