Following the pain

A correlation analysis between chest pain and cardiovascular disease in hospitalized adults

Malcolm Smith Fraser

11/14/20

IDS 702 – Modeling and Representation of Data

Contents

Summary

This analysis hopes to quantify the scale to which chest pain, exercise induced, and non-exercise induced, is an indicator of underlying cardiovascular pathologies. It will also explore what other predictors, when looked at in conjunction with heart disease, increase the correlation between the chest pain a patient reports and a true diagnosis of cardiovascular disease.

The regression uses patient data from the Cleveland clinic foundation. The response variable, a patient's angiographic heart disease status, is binary – referring to <50% arterial narrowing as a non-diagnosis, and >50% arterial narrowing as a diagnosis. For the predictive model, a logistic regression is used.

Introduction

For years now, cardiovascular diseases have been recognized as the leading causes of death both within the United States and globally. While there have been significant strides made towards developing metrics of identifying the progression of heart disease in individuals, the only predictor that is immediately recognized by a patient, without any tests, is their chest pain.

In exploring the correlation between chest pain and cardiovascular disease, I hope that an optimal course of action can be established for patients who feel experience chest pain. The goal of this analysis is not to find a single model that best predicts a patient's heart disease status, rather it's to determine what predictors, when observed in conjunction with chest pain are significant indicators of heart disease. Through the course of this analysis, I hope to come to find evidence-based answers to the following questions.

1. *Is chest pain an accurate signal of heart disease? What if it is exercise induced?*
2. Is the persistence of chest pain more significant if a patient has other confounding factors?
3. How do these effects compare for both men and women?

Data

The data was obtained from the UCI Machine Learning Repository[1]. Once cleaned, the data consists of 14 attributed and 297 observations. The proportion of participants with heart disease is 0.46. The response variable, disease status by vessel diameter narrowing, originally consisted of 5 levels, however for the scope of this analysis, all the levels indicating any progression of cardiovascular disease are grouped into a single level. Doing so allows us to focus simply on attempting to distinguish the presence of cardiovascular disease. The other variables are listed in *figure 3* below.

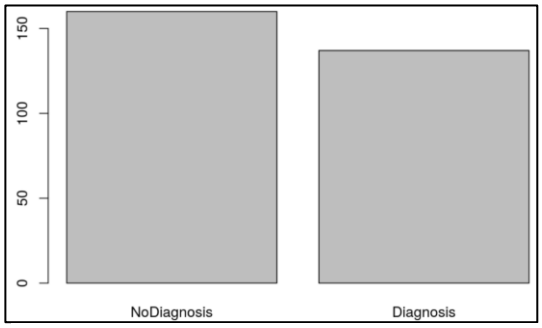

Figure 1. Distribution of the response variable across all observations.

Exploring the factor variables

Conditional probability tables were used to explore the relationship between the factor variables, and the response. The proportion of participants with a cardiovascular disease diagnosis appears to be lower for participants who experience regular chest pain and higher for those experiencing exercise

typical angina:		atypical angina:		non-anginal pain:		asymptomatic:	
NoDiagnosis	Diagnosis	NoDiagnosis	Diagnosis	NoDiagnosis	Diagnosis	NoDiagnosis	Diagnosis
0.6957	0.3043	0.8163	0.1837	0.7831	0.2169	0.2746	0.7254

Figure 2. Proportion of participants diagnosed with heart disease within the various chest pain categories. Significantly grater proportion of asymptomatic participants have a positive diagnosis.

induced chest pain. These relationships appear to be exacerbated when looking at the relationship between chest pain and the disease status given the status of exercise induced chest pain. For example, for asymptomatic participants that do not experience exercise induced chest pain, the proportion of positive to negative diagnosis are nearly equivalent. However, among asymptomatic (patients without regular chest pain) participants who do experience exercise induced chest pain, the proportion of positive diagnosis is 0.87. As such it may be worthwhile to explore modeling with the interaction between regular and exercise induced chest pain.

For both men and women, the proportion of participants with cardiovascular disease appears to be lower for participants who experience regular chest pain. For asymptomatic participants, the proportion of diagnosed male participants is considerably higher than non-diagnosed, while the proportion of non-diagnosed and diagnosed female participants are nearly equivalent. As such it may be worthwhile to explore the interaction between *sex* and chest pain.

Attribute	Description
Age	Continuous
Sex	Binary
Chest Pain (cp)	4 levels: asymptomatic, typical angina, atypical angina, non-anginal
Resting blood pressure (trestbps)	Continuous (mmHg)
Cholesterol levels (chol)	Continuous (mg/dl)
Fasting blood sugar (fbs)	(>120 mg/dl) 1=true, 0=false
Resting ecg (restecg)	Categorical: normal, ST-T wave abnormality, LV hypertrophy
Maximum HR achieved (thalach)	Continuous
Exercise induced angina (exang)	Binary
ST interval depression (oldpeak)	Continuous
Slope of ST interval (slope)	Categorical: upsloping, flat, downsloping
Number of blood vessels (ca)	Discrete (0-3)
Thallium stress test result (thal)	Categorical: normal, fixed defect, reversable defect
[Resp] Angiographic disease status (num)	0 : <50% diameter narrowing 1 : >50% diameter narrowing

Figure 3. Description of the variables found in the Cleveland dataset.

Thal, *slope*, and *ca* appear to have interaction effects with chest pain, however upon closer inspection these attributed lack sufficient observations at every level to make any meaningful conclusion about their interaction effects.

Exploring the continuous variables

Boxplots were used to explore the relationship between the continuous variables, and the response. The only attributes that appear to have an interaction effect with chest pain are *cholesterol* and *oldpeak*. Other interactions in the data appear to exist between *thal* and *age*, and *trestbps* and *sex*.

Modelling

When evaluating models and interpreting the results, the primary metric of interest was sensitivity. Sensitivity is the proportion of true positives to the total number of positive individuals. Since the goal of this analysis is to maximize the correlation between chest pain and heart disease, so as to be able to more efficiently recommend diseased participants the appropriate tests to confirm

asymptomatic: Male

NoDiagnosis	Diagnosis
0.2059	0.7941

asymptomatic: Female

NoDiagnosis	Diagnosis
0.45	0.55

Figure 4. Comparing the diagnosis proportions of asymptomatic male and female participants.

diagnosis, it makes sense to maximize the models ability to predict true positives – and focus less on the model's ability to distinguish true negatives which can be weeded out later.

Is chest pain an accurate signal of heart disease? What if it is exercise induced?

As a baseline, models where fit with chest pain and exercise induced chest pain as the only predictors.

$$baseline\_a = glm(num \sim relevel(cp, ref = "asymptomatic"), data = cleveland, family = 'binomial')$$

$$baseline\_b = glm(num \sim exang, data = cleveland, family = 'binomial')$$

The model achieves a sensitivity of 0.75, a specificity of 0.75, and an accuracy of 0.75. Exercise induced chest pain alone has a significantly lower sensitivity of 0.54, with similar specificity and accuracy. This tells us that exercise induced chest pain is not as closely correlated to cardiovascular disease as regular chest pain.

When looking at the model with both exercise-induced chest pain and non-exercised-induced chest pain, along with the interaction between those two predictors, the model performs identical to the model with chest pain alone.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.9712	0.188	5.165	2.4e-07
relevel(cp, ref = "asymptomatic")typical angina	-1.798	0.4906	-3.664	0.0002479
relevel(cp, ref = "asymptomatic")atypical angina	-2.463	0.4141	-5.948	2.719e-09
relevel(cp, ref = "asymptomatic")non-anginal pain	-2.255	0.326	-6.917	4.604e-12

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	409.9 on 296 degrees of freedom
Residual deviance:	328.8 on 293 degrees of freedom

Figure 5. Regression summary for the model with regular chest pain (cp) as the only predictor. All coefficients and the intercept are significant at an alpha level <.05.

The model with cheat pain alone shows that every level of chest pain is significant at a significance level of >.001. When compared to the chest pain baseline of asymptomatic, the odds of a patient with typical angina having a positive heart disease diagnosis is 83.43% lower, for participants with atypical angina it is 91.48% lower, and for participants with non-anginal pain it is 89.51% lower. These results are very counter intuitive as they signal that chest pain among these participants is negatively correlated with cardiovascular disease when compared to having no chest pain at all (asymptomatic).

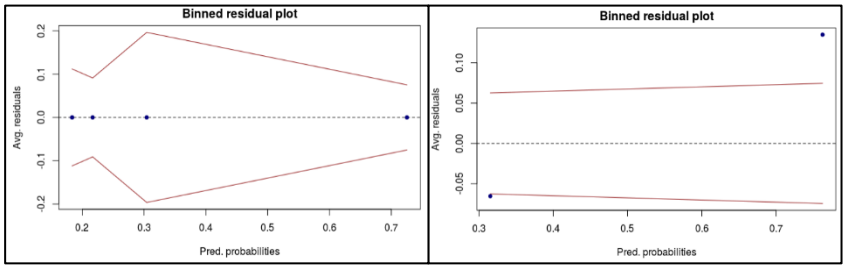

Figure 6. Binned residual plots for the model with regular chest pain (left) and the model with exercise induced chest pain (right). Chest pain model is a significantly better fit with all points lying at 0, within the plot bounds.

Is the persistence of chest pain more significant if a patient has other confounding factors?

Many of the other predictors have no effect of the model's predictive power, however, *thalach* – maximum (hear rate achieved), *oldpeak* (a measure of ST interval depression during exercise), *ca* (number of major blood vessels), and *thal* (thallium stress test results) all have an effect of the predicted outcomes when modeled with chest pain. Since the overall goal of this analysis is to find predictors

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.7953	0.1981	4.016	5.926e-05
relevel(cp, ref = "asymptomatic")typical angina	-1.493	0.5159	-2.894	0.003808
relevel(cp, ref = "asymptomatic")atypical angina	-1.962	0.4355	-4.504	6.665e-06
relevel(cp, ref = "asymptomatic")non-anginal pain	-1.982	0.341	-5.812	6.175e-09
thalach	-0.03347	0.007086	-4.724	2.314e-06

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	409.9 on 296 degrees of freedom
Residual deviance:	303.7 on 292 degrees of freedom

Figure 7. Regression summary for model with regular chest pain (cp) and thalach. The intercept and all predictors are significant at an alpha >.05.

that, along with chest pain, can best identify heart disease I will only be interpreting the models that saw changes in the predictive metrics. The predictors *thal* and *ca* are generated from fairly rigorous procedures used to diagnose specific heart defects and heart disease symptoms, as such it doesn't make sense to include them an early stage model geared towards helping guide future tests.

$$model\_1 = glm(num \sim relevel(cp, ref = "asymptomatic") + thalach, data = cleveland, family = 'binomial')$$

When *thalach* is included in the model with chest pain, we see an improved sensitivity of 0.77, a specificity of 0.78, and an accuracy of 0.77. Thus, in terms of predictive power, heart rate and chest pain together form a model that generates predictions that better distinguish between diseased and non-diseased participants. All predictors are significant at a significance level of 0. When compared the baseline of an asymptomatic patient with average *thalach*, the odds of a patient experiencing typical angina having a positive heart disease diagnosis is 77.53% lower, for participants experiencing atypical angina 85.93% lower, and for participants with non-anginal pain 86.22% lower. Also, for every one-unit increase in *thalach*, the odds of a patient having a positive diagnosis decreases by 3.29%. While these trend s are consistent with what was seen when modeling chest pain alone, the magnitude of the effect that the various forms of chest pain have does appear to have decreased at every level. Modeling with *thalach* makes the chest pain predictors weaker in terms of how much they decrease the odds of having a heart disease diagnosis, and thus not more significant of predictors.

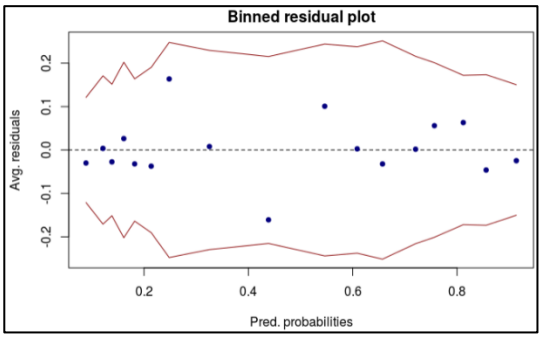

Figure 8. Binned residual plot for the model with chest pain (cp) and thalach.

$$model\_2 = glm(num \sim relevel(cp, ref = \text{asymptomatic}) * oldpeak, data = cleveland, family = 'binomial')$$

When *oldpeak* and its interaction with chest pain is included in the model, we see a decrease in sensitivity to 0.65, a specificity improvement of 0.83, and a similar accuracy of 0.75. In terms of predictive power, this model performs is worse than the baseline model at identifying chest pain since our primary metric is specificity. However, in terms of overall fit to the data the model does have an improved the deviance by 42.66. As in previous models, the coefficients for chest pain are all significant and signal similar trends. When compared to the baseline of an asymptomatic patient with average *oldpeak*, a one unit increase in *oldpeak* is correlated with an increased odds of heart disease diagnosis of 175%, however, given that the

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.9891	0.2176	4.544	5.511e-06
relevel(cp, ref = "asymptomatic")typical angina	-1.83	0.5229	-3.501	0.0004639
relevel(cp, ref = "asymptomatic")atypical angina	-1.951	0.5932	-3.288	0.001009
relevel(cp, ref = "asymptomatic")non-anginal pain	-2.276	0.3719	-6.12	9.377e-10
oldpeak	1.014	0.2413	4.202	2.648e-05
relevel(cp, ref = "asymptomatic")typical angina:oldpeak	-0.9713	0.4694	-2.069	0.03852
relevel(cp, ref = "asymptomatic")atypical angina:oldpeak	-0.2326	0.699	-0.3327	0.7394
relevel(cp, ref = "asymptomatic")non-anginal pain:oldpeak	0.08111	0.3948	0.2055	0.8372

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	409.9 on 296 degrees of freedom
Residual deviance:	286.1 on 289 degrees of freedom

Figure 10.Regression summary for t… on between chest pain and oldpeak. All… ak and atypical angina, an non-anginal p…

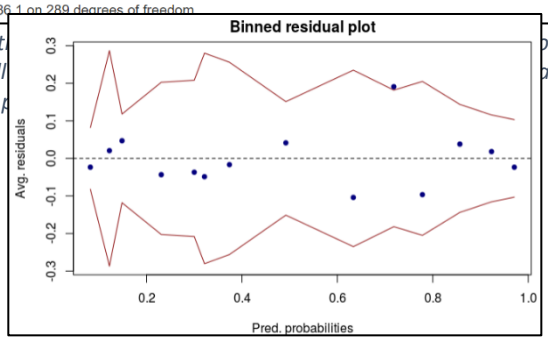

Figure 9. Binned residual plot for the above model.

patient is experiencing typical angina that increase in diagnosis odds is only 4.35%. As with *thalach*, modeling with *oldpeak* also decreases the magnitude of the effect that the chest pain predictors have on the odds of a heart disease diagnosis.

How do these effects compare for both men and women?

Exploring how sex effects the correlation between chest pain and cardiovascular disease does not find any significant results. It was found that the proportion of asymptomatic women diagnosed with heart disease is much higher than asymptomatic women. This, however, is not surprising. Chest pain in women is often not a heart attack symptom and it reasonable to assume that this correlation would be the same for signaling the progression of heart disease.

When modeling with the interaction between sex and chest pain, compared the baseline of asymptomatic and male, all the chest pain levels are significant at α=.001. However, given that sex is female, when compared to the same baseline, none of the chest pains are significant. Thus, the correlation between chest pain and cardiovascular disease is not equal for both men and women.

When modeling with chest pain and *thalach*, compared to the baseline of asymptomatic and male at the average *thalach* value, all the chest pain levels are significant at α=.05 and *thalach* is significant at α=.001. However, given the sex is female, compared to the same baseline, the only significant predictor is the asymptomatic and non-anginal chest pain predictors. This, the observed relationship that chest pain and *thalach* have with cardiovascular disease is not the same for both men and women.

Conclusions

The results of this analysis were very different than what was expected. While chest pain is correlated with cardiovascular disease, the correlation is negative. As such, when answering the first question of interest – *Is chest pain an accurate signal of heart disease?* – the answer is no. While models with chest pain as a predictor do prove more effective in identifying cardiovascular disease there the analysis showed evidence that the chest pain is in no way "signaling" said disease status. Addressing the second part of the question – *What if it is exercise induced?* – the answer is yes, when modelling with exercise induced angina as a predictor the attribute is positively correlated with increased odds of a positive disease diagnosis. However, it must be noted that at best, models that included exercise induced angina performed at the same level of those with regular chest pain, and at worst performed significantly worse.

Addressing the second inferential question – *Is the persistence of chest pain more significant if a patient has other confounding factors?* – the answer is sort of… The original scope of this question assumed that chest pain was positively correlated with cardiovascular disease and would have evaluated the coefficient of the chest pain levels as other predictors were added to the model. However, as the initial assumption did not hold the evaluation of the question needed to change. Models with chest pain along with *thalach* (maximum heart rate achieved) improved the sensitivity of the model. This means that the model was better able to identify participants who did have cardiovascular disease. However, the notion that chest pain is "more significant" is not supported as the coefficients for the chest pain levels moved closer to zero when compared to the model without *thalach*.

Lastly, addressing the third inferential question – *How do these effects compare for both men and women?* – we have evidence that chest pain and cardiovascular disease are not correlated the same way for both men and women. Across all the models that include chest pain the different levels of chest pain appear to have less effect

on predicting the disease status in women than for men. This is a trend that we saw throughout EDA and one that does not surprise us given that chest pain, when referring to hear attacks is a less consistent side effect for women.

One limitation to this study is limited amount of data which prevented certain interactions from being fully explored as there were insufficient observations at certain levels. Another significant limitation in this analysis is the unclear selection criteria of the participants. This second limitation could have seriously affected the result of this analysis and could potentially have had an influence on the unexpected outcomes observed.

References

[1] DATA - https://archive.ics.uci.edu/ml/datasets/Heart+Disease

V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.