Project 2: Coronary Heart Disease

Tim Edinger, Max St. Clair, Victor Pan

Summary

This paper aims to analyze the core causes of coronary heart disease (CHD) through the usage of linear regression, k-means clustering (kMC), and k-nearest neighbor (kNN), then develop an accurate model to predict occurrences of CHD in patients. In order to conduct this analysis, data was first cleaned and imputed, then linear regression and kMC were used to determine the variables that were most correlated with CHD. Once this was completed, the variables with the most significant P-values from linear regression and clustering from kMC were used to complete a kNN regression analysis, developing a model that could be used to predict occurrences of CHD for any data set. From our linear regression and kMC analysis, we found that the variables 'age', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate' and 'glucose' were most statistically significant, which were then used in kNN. Our kNN model developed using these variables had an accuracy of roughly 0.8 when trialed using the test dataset.

Data

The data was provided to us in a largely cleaned state, with few missing values and quantitative variables in numerical (float) data type. This made the task of cleaning variables relatively easy. We were not concerned with the quantity of missing values nor

whether the occurrence of missing values was endogenous to the model, so we decided to simply drop the missing values. If we were concerned with the quantity of missing values, we could have imputed values for missing to the median, and marked with a dummy variable which values were imputed. Concerns with endogeneity of missing values would need to be addressed in the data collection process. But, we were able to simply drop missing values.

The target value for our predictions was the variable TenYearCHD, the 10 year risk of developing coronary heart disease. The predictor variables are the following:

- sex : the recorded sex of the observations with 1 denoting an participant coded as male
- age : Age at the time of medical examination in years.
- education: A categorical variable of the participants education, with the levels: Some high school (1), high school/GED (2), some college/vocational school (3), college (4)
- currentSmoker: Current cigarette smoking at the time of examinations
- cigsPerDay: Number of cigarettes smoked each day
- BPmeds: Use of Anti-hypertensive medication at exam
- prevalentStroke: Prevalent Stroke (0 = free of disease)
- prevalentHyp: Prevalent Hypertensive. Subject was defined as hypertensive if treated
- diabetes: Diabetic according to criteria of first exam treated
- totChol: Total cholesterol (mg/dL)

- sysBP: Systolic Blood Pressure (mmHg)
- diaBP: Diastolic blood pressure (mmHg)
- BMI: Body Mass Index, weight (kg)/height (m)^2
- heartRate: Heart rate (beats/minute)
- glucose: Blood glucose level (mg/dL)

To process the variables for linear regression, first we wanted to include log values for the non-zero quantitative variables. Specifically, ['age', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose'] were used. cigsPerDay is also a quantitative variable, but includes zero values for non-smokers so it cannot be used for log values. In addition to taking log values, we wanted to incorporate interaction terms into the model. The variable sex was chosen to interact with the quantitative variables. To create these log and interaction terms, we wrote a loop that cycles through the quantitative variables.

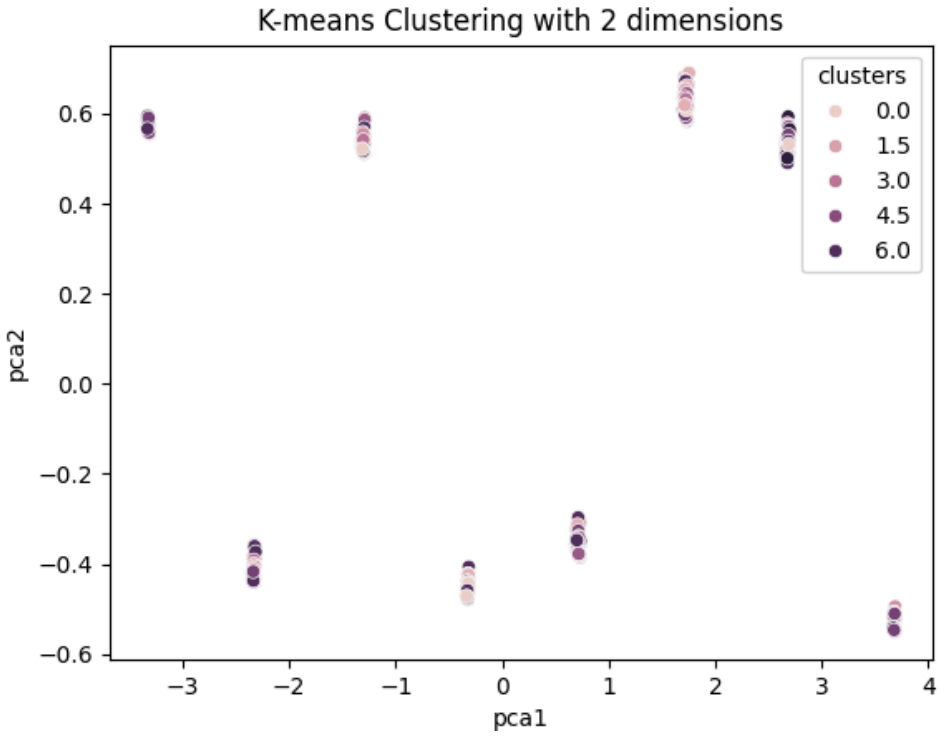
```
vars = ['age', 'totChol', 'sysBP',  'diaBP',  'BMI',  'heartRate',  'glucose'] # quantitative vars we want to log and interact with gender
for i in vars:
  df[i+'_log'] = np.log(df[i]) # take the log form of variables
  df[i+'_sex'] = df[i]*df['sex'] # create an interaction term with sex

# same procedure for test set
for i in vars:
  df_test[i+'_log'] = np.log(df_test[i])
  df_test[i+'_sex'] = df_test[i]*df_test['sex']
df.head()
```

Results

For the implementation of k-means clustering the training dataset was used to develop groupings for the data set that were iterated upon until clear clusters were created. In order to accomplish this we began iterating with k-means clustering , assuming 2 clusters in the data set, 300 max iterations, and 10 initial centroid seeds. The

number of clusters and the dataset considered was changed until an optimal configuration was reached, one that had clear differences between clusters with and without TenYearCHD. In order to analyze this effectiveness, a number of measures were used including visualization via principal component analysis (PCA), scree plots, and pandas groupby means. PCA was used as a method to allow for 2D visualization of a greater than two dimensional dataset and as 6 features are used in this analysis, PCA is required or traditional visualization would be impossible.



The graph above shows one such iteration of PCA visualization, with clusters based around +/- 0.5 due to the binary nature of the variables ‘sex’ and ‘TenYearCHD. Because of this relationship clusters were mainly grouped around male/female and CHD/no CHD, with other trends becoming present as the number of clusters increased. Iteration and scree plots showed that an optimal number of clusters was 8, as trends for

variables relating to CHD were most visible. These initial conditions produced the following result:

clusters	cigsPerDay	sysBP	glucose	age	sex	prevalentStroke	TenYearCHD
0	0.034589	0.313144	0.124212	0.756642	0.0	0.012862	0.0
1	0.381736	0.205315	0.108146	0.317855	1.0	0.000000	0.0
2	0.206461	0.270514	0.135968	0.566172	1.0	0.016598	1.0
3	0.075644	0.306618	0.136813	0.610440	0.0	0.010929	1.0
4	0.057384	0.233948	0.114469	0.489318	0.0	0.002114	0.0
5	0.097324	0.238971	0.120974	0.685165	1.0	0.012085	0.0
6	0.119225	0.170384	0.108221	0.236646	0.0	0.001859	0.0
7	0.025735	0.214895	0.116214	0.277283	1.0	0.003676	0.0

The table above shows that two main clusters with CHD were found, one for male and one for female, each with a number of interesting trends among other variables. Males with CHD (in cluster 2), denoted by the variables ‘sex’ and ‘TenYearCHD’ both being equal to 1, had above average values of ‘prevalentStroke’, ‘glucose’, ’age’, ‘sysBP’, and ‘cigsPerDay’ signifying that these variables are good predictors of coronary heart disease. Females with CHD (in cluster 3) had similar trends as compared to the males, though cigsPerDay does not appear to be an accurate predictor of CHD in this case. Overall females tended to have higher values of sysBP, glucose, and age in clusters with CHD when compared to males. In general k-means showed that, across gender, systolic blood pressure, blood glucose levels, presence of stroke, and age were the best predictors for CHD. This relationship was found both when all variables were included in analysis, and in the testing database, signifying that this clustering is somewhat significant.

While kMC is helpful for finding trends that may otherwise be missed in the data, we need to implement a supervised model to achieve an r^2 for the project. One of the ways we did this was using a linear model that incorporated log values and interaction terms. The variable sex was chosen to interact with the quantitative variables. This choice allows the model to differentiate between effects of the quantitative variables in men and women. Take this simplified model as an example.

$$\hat{y} = \beta_0 + \beta_1 * sex + \beta_2 * sysBP + \beta_3 * sex * sysBP + e$$

In this case, beta2 captures the effect of systolic blood pressure on women (when sex = 0), and the variable beta3 captures the additional effect of blood pressure on men. So, for each 1 mmHg increase in a male's blood pressure, their chance of developing CHD increases by beta2 + beta3. Sex makes sense as an interaction term because it allows these effects to be isolated to men and women, whose bodies presumably react to these factors somewhat differently. The sign of the coefficient on the interaction term shows whether men or women are more sensitive to the respective variable. A positive interaction term would imply that that factor raises male risk of CHD more than female risk.

We used statsmodel.api to run the MLR because we were interested in the r^2 on the training set, coefficients and p-values. The model regressed TenYearCHD on all of the given variables plus the log values and interaction terms we calculated. Here are the results:

OLS Regression Results

Dep. Variable:	TenYearCHD	R-squared:	0.116
Model:	OLS	Adj. R-squared:	0.107
Method:	Least Squares	F-statistic:	12.34
Date:	Tue, 23 Apr 2024	Prob (F-statistic):	2.51e-54
Time:	02:05:25	Log-Likelihood:	-931.24
No. Observations:	2744	AIC:	1922.
Df Residuals:	2714	BIC:	2100.
Df Model:	29		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.6241	2.611	3.304	0.001	3.505	13.743
sex	-0.4250	0.157	-2.708	0.007	-0.733	-0.117
age	0.0158	0.009	1.667	0.096	-0.003	0.034
education	-0.0017	0.007	-0.258	0.797	-0.015	0.011
currentSmoker	-0.0214	0.021	-1.020	0.308	-0.063	0.020
cigsPerDay	0.0028	0.001	3.125	0.002	0.001	0.005
BPMeds	0.0676	0.041	1.664	0.096	-0.012	0.147
prevalentStroke	0.0940	0.084	1.115	0.265	-0.071	0.259
prevalentHyp	0.0174	0.020	0.867	0.386	-0.022	0.057
diabetes	0.0141	0.052	0.269	0.788	-0.089	0.117
totChol	0.0030	0.001	2.913	0.004	0.001	0.005
sysBP	0.0045	0.003	1.411	0.158	-0.002	0.011
diaBP	0.0057	0.007	0.834	0.404	-0.008	0.019
BMI	0.0172	0.012	1.374	0.169	-0.007	0.042
heartRate	-0.0025	0.005	-0.516	0.606	-0.012	0.007
glucose	0.0027	0.001	2.584	0.010	0.001	0.005
age_log	-0.4695	0.468	-1.003	0.316	-1.387	0.448
age_sex	0.0031	0.002	1.771	0.077	-0.000	0.007
totChol_log	-0.6860	0.257	-2.674	0.008	-1.189	-0.183
totChol_sex	0.0006	0.000	1.937	0.053	-7.74e-06	0.001
sysBP_log	-0.3862	0.448	-0.863	0.388	-1.264	0.491
sysBP_sex	0.0016	0.001	1.471	0.141	-0.001	0.004
diaBP_log	-0.6054	0.565	-1.071	0.284	-1.714	0.503
diaBP_sex	0.0004	0.002	0.209	0.834	-0.003	0.004
BMI_log	-0.4132	0.339	-1.218	0.223	-1.079	0.252
BMI_sex	-0.0041	0.004	-1.112	0.266	-0.011	0.003
heartRate_log	0.1719	0.374	0.460	0.645	-0.561	0.905
heartRate_sex	0.0005	0.001	0.441	0.659	-0.002	0.003
glucose_log	-0.1951	0.102	-1.908	0.056	-0.396	0.005
glucose_sex	0.0002	0.001	0.348	0.728	-0.001	0.001

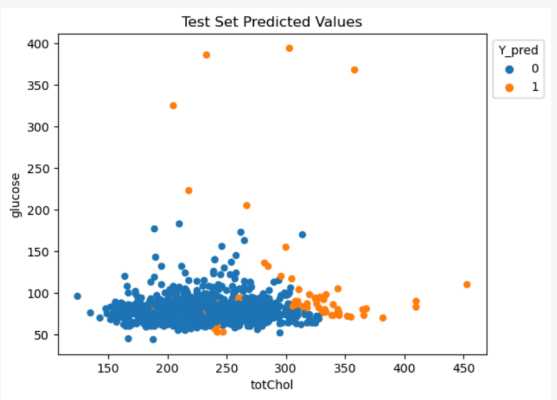
Omnibus:	735.837	Durbin-Watson:	1.977
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1493.022
Skew:	1.624	Prob(JB):	0.00
Kurtosis:	4.583	Cond. No.	1.41e+05

Given the sheer number of variables considered in the model, it is not a surprise that many coefficients were not significant at the p = 0.05 level. This model is not particularly carefully specified, and leads to concerns of multicollinearity. For example, there is good intuition that systolic and diastolic blood pressure are closely related, making either of their coefficients statistically unreliable.

Yet, our primary goal is not to discover and quantify the effects of these variables on risk of CHD. Rather, we are trying to create a predictive model for future cases of

CHD given the model variables. To evaluate the predictive efficiency of this model, we used the test set. With the test set values, we predicted yhat, the values for TenYearCHD. To do this, we calculated the sum of squared errors by summing the residual of predicted values against actual values. Then we calculated the total square errors by doing the same but simply using the mean value as the predictor. r^2 is calculated as 1 - SSE/TSS. The model gave an r^2 of 0.102 on the test set. In other words, the model explained 10.2% of the variance in outcomes for developing CHD.

Then, to develop the kNN model, we used insights from kMC and MLN to train a kNN on the dataset fhs_train.csv and applied to the dataset fhs_test.csv to predict the risk of a 10-year coronary heart disease event (TenYearCHD) using total cholesterol level (totChol) and glucose (glucose) as the input variables. The k value for the kNN algorithm is set to 7, meaning that the model considers the 7 nearest neighbors to make a prediction, and the Euclidean distance metric has been used to calculate the distance between data points.



The scatter plots above illustrate the actual and predicted values of the TenYearCHD in the test set using total cholesterol levels (totChol) and glucose (glucose) as the input variables. The right scatter plot displays the actual TenYearCHD against the input variables totChol and glucose, with blue dots representing individuals with low risk of TenYearCHD and orange dots representing individuals with a high risk of TenYearCHD. The left scatter plot displays the predicted TenYearCHD values, again with blue dots for predictions of low risk of TenYearCHD and orange dots for predictions of high risk of TenYearCHD.

From the scatter plots, observe:

The actual values plot (right) displays a distribution of cholesterol levels and glucose levels amongst individuals with high and low TenYearCHD, with no clear boundary separating the two groups. However, individuals with very high glucose levels seem to have a very high likelihood of having a high risk of TenYearCHD. Also, individuals with extremely high cholesterol levels seem to have a low risk of TenYearCHD which is interesting and might indicate that there might be other factors in play that affect the risk of an individual having a high TenYearCHD.

The predicted values plot (left) displays how the kNN model classifies the test set data points. The model seems to predict more high TenYearCHD values at higher cholesterol levels, which aligns with the general consensus that higher cholesterol levels can increase the risk of heart disease.

Comparing the actual and predicted plots, we can see that our model has captured some of the patterns in the data, but there are also areas where the predictions are

completely off the mark. This is somewhat expected, though, as the kNN algorithm is a simple model and may not capture all of the complex relationships in the data.

The accuracy of this model, which is a measure of how many predictions were correct out of all predictions made, was calculated using the .score function and wasn't too bad at a value of 0.806282722. To improve this model's performance it is possible to add additional relevant features as inputs to this model.

CONCLUSION

In conclusion, our analysis aimed to predict the 10-year risk of CHD using various predictor variables. We implemented a multiple linear regression model to predict the risk of CHD based on various variables, log values, and interaction terms. Based on the results of the linear regression model, we utilized the KNN algorithm and k-means clustering to explore relationships between total cholesterol level, glucose level, and the risk of CHD.

The KNN model showed some success in predicting the risk of CHD based on total cholesterol and glucose levels. However, it also highlighted areas where the predictions were off the mark, indicating the limitations of the KNN algorithm in capturing complex relationships in the data. The kMC analysis provided insights into the predictors of CHD, with variables such as systolic blood pressure, blood glucose levels, presence of stroke, and age emerging as significant predictors. This analysis helped identify trends in the data that may have been overlooked in traditional regression analysis. The MLR model, while providing predictive power, also had limitations. Many

coefficients were not significant at the 0.05 level, indicating potential multicollinearity issues. This suggests that some variables may be redundant or highly correlated, leading to unreliable coefficient estimates.

To address these issues, future analyses could explore other modeling techniques, such as LASSO, which can help identify and select the most relevant variables for predicting CHD. This could improve the predictive accuracy of the model and help mitigate issues of multicollinearity. Overall, our analysis provides valuable insights into the predictors of CHD and highlights the importance of using a variety of analytical techniques to explore complex datasets. Further research and refinement of models could lead to better prediction and understanding of CHD risk factors.