Analysis of Framingham Heart Data

Advanced Data Analysis Final Project

Prof. Hammou Elbarmi

Tay Hui Chiang (ht2490), Thuy Linh Nguyen (tn2382)

Contents:

Summary

Motivation for the Project

The purpose of the Framingham Heart data set was to determine the etiology of cardiovascular diseases in the community of Framingham, Massachusetts. We decided to perform our statistical analysis on this data set because we were interested in finding out whether the risk factors that were investigated in the study could actually affect the life expectancy of a patient, and the patient's cause of death.

We will consider two cases based on the known statuses of patients at the end of the study. In the first one the outcome variable indicates a patient's status at the end of the study, and in the second one the outcome variable indicates a patient's cause of death. We would like to identify potential risk factors and their joint effects on the odds of death of cancer, coronary heart disease, cerebral vascular disease and others when compared to the odds of being alive (first case), as well as the odds of death of coronary heart disease, cerebral vascular disease and others when compared to the odds of death of cancer (second case). Likely influential factors include, among others, blood pressure, height, weight, cholesterol level and smoking status.

Summary of Results

Ideally, we would have liked to perform a survival analysis study, however this was not possible because the censored timings for patients who did not die during the duration of the study was not available. Due to the categorical outcome variable with multiple factors, we performed multinomial logistic regression on the Framingham Heart dataset.

The results of the analysis indicate that, in the first case, Weight and Diastolic variables do not have an effect on the odds ratios, and therefore can be dropped from the model that predicts a patient's status at the end of the study. Sex, AgeAtStart, Systolic and Smoking turn out to be very significant in predicting the statuses of patients.

In the second case, we conclude Smoking and Diastolic can be dropped from the full model that predicts the cause of death. In the model without those variables, Height, Weight, Systolic and MRW are found to be significant predictors for causes of death.

Data

In this section, we will be detailing the Framingham Heart Data that was analysed for this project by looking at the variables that were provided. We will also do some exploratory data analysis to find out visually if there were any possible trends we could look out for, or any correlation between predictor variables.

Description of Data

The data set consisted of 5210 patients (observations), and 17 variables, listed below:

1. Status - Categorical variable with 2 factors, ‘Dead’ and ‘Alive’, describing the status of the patient by the end of the study.
2. DeathCause - Categorical variable with 4 factors, ‘Cancer’, ‘Cerebral Vascular Disease’, ‘Coronary Heart Disease’, ‘Other’ and ‘Unknown’. For patients who are alive, this entry is filled with NA.
3. AgeCHDdiag - Numerical variable describing the age at which the patient was diagnosed with Coronary Heart Disease. Most of the entries are NA.
4. Sex - Categorical variable with 2 factors, ‘Male’ and ‘Female’.
5. AgeAtStart - Numerical variable for the age of the patient at the start of the study.
6. Height - Numerical variable.
7. Weight - Numerical variable.
8. Diastolic - Numerical variable for blood pressure.
9. Systolic - Numerical variable for blood pressure.
10. Metropolitan Relative Weight (MRW) - Numerical variable for a measure of body fat based on height and weight.
11. Smoking - Numerical variable for number of frequency of cigarettes smoked
12. AgeAtDeath - Numerical variable for age at which the patient died. For patients who are alive, this entry is NA.
13. Cholesterol - Numerical variable for cholesterol level in blood.
14. Chol_Status - Categorical variable with 3 factors based on cholesterol, ‘Borderline’, ‘Desirable’ and ‘High’.
15. BP_Status - Categorical variable with 3 factors based on diastolic and systolic blood pressure, ‘High’, ‘Normal’ and ‘Optimal’.

16. Weight_Status - Categorical variable with 3 factors based on MRW, ‘Normal’, ‘Overweight’ and ‘Underweight’.
17. Smoking_Status - Categorical variable with 5 factors based on smoking, ‘Heavy’, ‘Light’, ‘Moderate’, ‘Non-Smoker’, ‘Very Heavy’.

For our analysis, we did not include variables 14-17 because the information in these variables were already present as numerical variables with more precision. We also did not include ‘AgeCHDdiag’ and ‘AgeAtDeath’. Since we wanted to look at how various risk factors affected whether the patient was alive at dead at the end of the study, and if they died what caused it, we then merged ‘Status’ and ‘DeathCause’ to form a new ‘Status’ categorical variable with factors ‘Cancer’, ‘Cerebral Vascular Disease’, ‘Coronary Heart Disease’, ‘Other’ and ‘Alive’. Lastly, we removed any patients with NA entries in columns such as cholesterol, as well as those who had died with cause ‘Unknown’.

Exploratory Data Analysis

We performed exploratory data analysis by first creating a correlation matrix for the 8 numerical predictor variables (AgeAtStart, Height, Weight, Diastolic, Systolic, MRW, Smoking and Cholesterol) to see if there are any possible collinearities in the data. This is presented in Table 1 in the Appendix.

We can see in Table 1 that most of these predictors seem to be independent of each other, with the exception of Systolic and Diastolic with an R^2 coefficient of 0.797, and MRW and Weight with an R^2 coefficient of 0.765.

We also make box and whisker plots of various numerical predictors against the patient status to see if they may be significant. For example, in Figure 1a in the Appendix, we see that patients who died of Cerebral Vascular Disease or Coronary Heart Disease during the duration of the study also tended to have higher Diastolic blood pressures. A similar trend can be seen for Cholesterol level, in Figure 1b.

Lastly, we made a bar plot for the frequency of each patient’s status (Figure 1c). We can see that the majority of patients were alive at the end of the study, which implies that the data is

unbalanced. Thus it would also be worth focusing on only the dead patients and the causes of death further on in our analysis.

Analysis of Results

In this section, we perform a multinomial logistic regression with two different predictors, namely the status of all patients (Alive or Dead due to Cancer etc.), and the causes of deaths for patients who are dead. Both logistic regressions will be performed initially on the full set of predictor variables. The full model can be expressed as follows:

status ~ Sex + AgeAtStart + Height + Weight + Diastolic + Systolic + MRW + Smoking + Cholesterol

Regression with Status of All Patients

We first note that there are 5 possibilities for a patient's status at the end of the study: 'Alive', 'Cancer', 'Cerebral Vascular Disease', 'Coronary Heart Disease', and 'Other'. The model that we used assumes the status of 'Alive' as a base case, and uses the logarithmic odds of each other status occurring against the 'Alive' status as a response. For example, if we looked at 'Cancer', then the model would be,

$$\log\left(\frac{Probability\ that\ patient\ died\ by\ Cancer}{Probability\ that\ patient\ is\ Alive}\right) = \beta_{0,Cancer} + \beta_{1,Cancer}X_1 + \cdots + \beta_{10,Cancer}X_{10}$$

In the above equation, the β's are coefficients that we estimate for each predictor variable and each patient status, and the X's are the predictor variable values. For example, X_1 represents the patient's sex, and X_{10} represents the patient's cholesterol level. The full table for coefficients can be found in Table 2a.

If we look at the categorical predictor sex, given that the response is either the patient being Alive, or a specific other status of interest, holding everything else fixed, a Male would have the odds of that status occurring increase by a multiplicative factor of $e^{\beta_{sex}}$, where β_{sex} is the coefficient for sex and that specific status in Table 2a. For the other numerical predictors, an increase in 1 unit of that predictor, holding everything else fixed, would see the odds of the

status occurring increase by a multiplicative factor of e^{β}, where β is the coefficient for the predictor variable and status of interest.

We can also observe the p values for each coefficient in Table 2b. We see that at a 95% level of significance, almost all the predictor variables have an effect on the odds of each status, with the exception of Weight, Diastolic, MRW and Cholesterol for certain statuses. However, we did anticipate this to occur from our exploratory data analysis as some of these variables are correlated with each other.

We now remove Weight and Diastolic from the model and re-run the regression on this reduced model. The model used is similar to the one above, except that we have 2 fewer predictors. The corresponding tables of coefficients and p values can be found in Tables 2c and 2d respectively. Upon performing a Likelihood Ratio Test, using the respective residual deviances of the reduced and full models, on a χ^2-distribution with 8 degrees of freedom, we see that the p-value for this test is 0.056. Hence, at a 95% level of significance, we may conclude that this reduced model could be a better fit in predicting the patient's status at the end of study.

However, as mentioned earlier, the data is unbalanced and the majority of patients are Alive at the end of the study. We now perform a similar analysis, this time with the response being the causes of death among patients who died during the course of the study.

Regression with Causes of Death

Let us now consider a different model, in which only dead patients are of interest. Removing the 'Alive' category from 'status' leaves us with 4 remaining possibilities for the response variable, which now indicates cause of death: 'Cancer', 'Cerebral Vascular Disease', 'Coronary Heart Disease', and 'Other'. We will start with the same covariates for the full model as before. We assume the reference group to be those patients who died of cancer. The reference group for the Sex variable are females. We interpret the slope coefficients as previously. The results of fitting the full model are included in Table 3a (coefficients) and Table 3b (p values).

Among others, Height, Weight and MRW variables have significant coefficient p values for all of the 3 levels of the outcome (since cancer is the reference group). Say for Height, we can

interpret the coefficients as follows: given that the response is either ‘Cerebral Vascular Disease’ or ‘Cancer’, if we increase Height by 1 while holding everything else fixed, the odds of the response is ‘Cerebral Vascular Disease’ change by a multiplicative factor equal to $e^{\beta_{\text{Height, Cerebral Vascular Disease}}} = e^{-0.14611} \approx 0.86$, i.e., roughly 14% decrease. Given that the response is either ‘Coronary Heart Disease’ or ‘Cancer’, if we increase Height by 1 while holding everything else fixed, the odds of the response is ‘Coronary Heart Disease’ change by a multiplicative factor equal to $e^{\beta_{\text{Coronary Heart Disease, Cancer}}} = e^{-0.14188} \approx 0.87$, i.e., roughly 13% decrease. Finally, given that the response is either ‘Other’ or ‘Cancer’, if we increase Height by 1 while holding everything else fixed, the odds of the response is ‘Other’ change by a multiplicative factor equal to $e^{\beta_{\text{Other, Cancer}}} = e^{-0.24215} \approx 0.78$, i.e., roughly 22% decrease.

It appears to be the case that the coefficients for Sex and Cholesterol variables for ‘Coronary Heart Disease’ outcome when these compared to the reference group ‘Cancer’ are significant, but not significant for ‘Cerebral Vascular Disease’ and ‘Other’ when these compared to the reference group ‘Cancer’. Moreover, the Smoking variable indicating the number of the frequency of cigarettes smoked is insignificant for all three outcomes when they are compared to the baseline. We can assume that smoking increases the risk of considered causes of death equally and is thus not needed to in our model. Again, we know from exploratory analysis that Diastolic and Systolic are correlated to each other. Hence, for the reduced model, we will drop both Smoking and Diastolic variables.

The slope coefficients and p-values are found in Table 3c and Table 3d. We see that the slope coefficients for Height, Weight, MRW and Systolic are all significant. The negative slope for Height indicates a decrease in the odds that the response is ‘Cerebral Vascular Disease’, ‘Coronary Heart Disease’ or ‘Other’ when compared with ‘Cancer’ if we increase Height and hold other variables constant. The same applies to the MRW variable. It seems to be the case that individuals with increased Weight or Systolic measures are more prone to die of ‘Cerebral Vascular Disease’ given that the response is that or ‘Cancer’, of ‘Coronary Heart Disease’ given that the response is that or ‘Cancer’, and of ‘Other’ diseases given that the response is that or ‘Cancer’. The Sex variable has a significant slope only for ‘Coronary Heart Disease’ outcome. AgeAtStart is only significant for ‘Cerebral Vascular Disease’, and Cholesterol is significant for ‘Coronary Heart Disease’ and ‘Other’.

We conduct ANOVA to compare the two fits. The null hypothesis states that we do not need the Smoking and Diastolic, and the alternative hypothesis is the full model with those variables included. We find that the Likelihood Ratio statistic is 8.46 and the corresponding p value is 0.2063. Hence, we cannot reject the null hypothesis at 5% significance level, and conclude that the reduced model is better.

Residual Diagnostics

We also assess the fit of our reduced models for both analyses based on the residuals of the models. To do this, we use a binned plot of residuals against expected values. This means that for each response category ('Alive', 'Cancer' etc), we create bins for the predicted probabilities. We then calculate the average residual and average predicted probabilities for each bin, and then plot the residuals against the probabilities. In the plots, we expect the average residuals to be approximately normally distributed about the x-axis, with no trend against the expected probabilities.

For the first analysis which includes 'Alive' as a patient status, we make a binned plot of the average residuals against expected values for 'Cancer'. We see that the residuals are largely symmetric about the x-axis. However, they seem to increase as the expected values increase, which would imply the presence of heteroscedasticity. A similar trend can be observed for the other response variables. While this is not ideal, homoscedasticity is not an assumption for logistic regression, and thus we should not have to modify our analysis to account for this. We also note that the grey lines in the plot indicate a confidence interval of 2 standard deviations. Our plot does correspond with this since a majority of the points lie between the lines, while a small minority lie outside them, which is what we expect.

Conclusion

We conclude that in both of the regressions considered, with statuses of all patients at the end of study and causes of death, dropping the variables with most insignificant slope coefficients improved the model fits. For the first case, Sex, AgeAtStart, Systolic and Smoking was found to have a strong predictive power for a patient's status. For the second case, Height, Weight, Systolic and MRW was found to have a strong predictive power for a patient's cause of death.

The variable Diastolic was highly correlated with Systolic, and therefore it had to be removed from the full model. Moreover, Height, Weight and MRW were correlated with each other, which resulted in some insignificant slope estimates. Finally, Smoking covariate turned out to be ineffective in predicting the causes of death since it is generally thought to be the cause of many various diseases.

Our results for the first analysis generally corroborate the popular consensus that males do tend to live shorter lives than females, and that factors such as blood pressure, smoking and body fat index are indeed risk factors for diseases.

For the second analysis, as mentioned previously, smoking is ineffective as a predictor for specific diseases. It is also interesting to note that for Coronary Heart disease, AgeAtStart is not a significant predictor, which implies that this disease could affect people of all ages, and that other risk factors based on health and lifestyle such as blood pressure and cholesterol are more significant in increasing the odds of occurrence of this disease over other diseases.

Appendix

Table 1. Correlation Matrix for Numerical Predictor Variables								
	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol
AgeAtStart	1.00000	-0.13316	0.09172	0.27808	0.38108	0.20426	-0.17092	0.28015
Height	-0.13316	1.00000	0.52540	-0.01083	-0.06823	-0.13013	0.28971	-0.07725
Weight	0.09172	0.52540	1.00000	0.32900	0.26280	0.76529	0.09149	0.07648
Diastolic	0.27808	-0.01083	0.32900	1.00000	0.79703	0.38651	-0.06695	0.18494
Systolic	0.38108	-0.06823	0.26280	0.79703	1.00000	0.36175	-0.09144	0.19992
MRW	0.20426	-0.13013	0.76529	0.38651	0.36175	1.00000	-0.12553	0.14054
Smoking	-0.17092	0.28971	0.09149	-0.06695	-0.09144	-0.12553	1.00000	-0.01411
Cholesterol	0.28015	-0.07725	0.07648	0.18494	0.19992	0.14054	-0.01411	1.00000

Figure 1a. Box Plot of Diastolic BP against Status

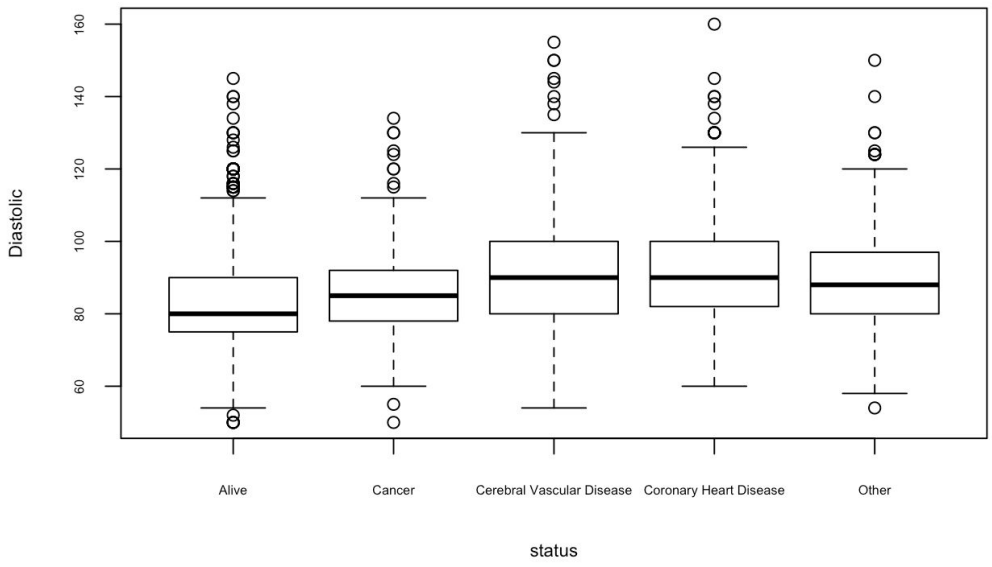

Figure 1b. Box Plot of Cholesterol against Status

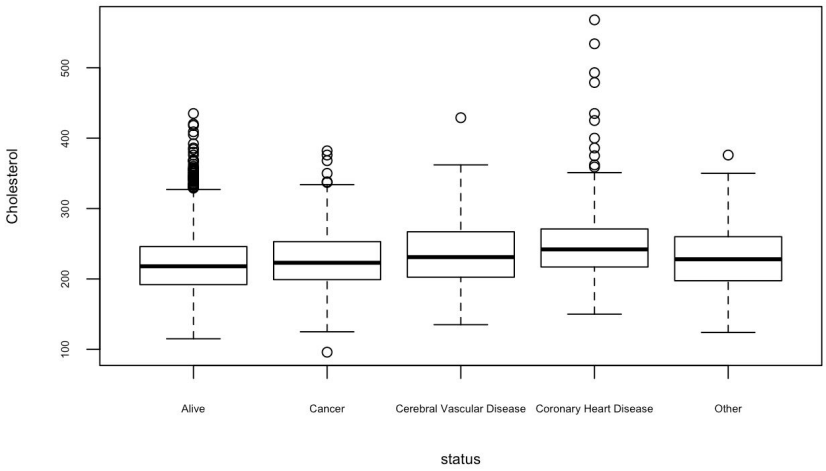

Figure 1c. Bar Plot of Patient Status

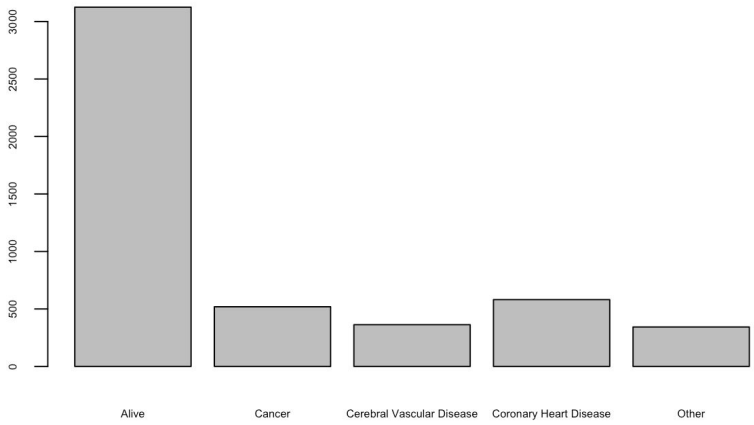

Figure 2a. Table of Coefficients for Full Model

	Intercept	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol
Cancer	-16.14737	0.43430	0.10331	0.13431	-0.02622	0.00025	0.00795	0.03248	0.03220	-0.00181
Cerebral Vascular Disease	-11.54257	0.51981	0.14678	-0.02414	0.00007	0.00113	0.02520	0.00130	0.03447	z-0.00081
Coronary Heart Disease	-12.10520	1.30115	0.11501	-0.02675	0.00341	0.01245	0.01780	0.00159	0.02762	0.00718
Other	-1.23380	0.49500	0.12023	-0.13728	0.01820	0.01887	0.00894	-0.02663	0.03049	-0.00195

Figure 2b. Table of p-Values for Full Model										
	Intercept	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol
Cancer	0.00000	0.00600	0.00000	0.00000	0.00000	0.96952	0.03058	0.00000	0.00000	0.12636
Cerebral Vascular Disease	0.00000	0.00641	0.00000	0.02022	0.99022	0.87485	0.00000	0.84553	0.00000	0.56132
Coronary Heart Disease	0.00000	0.00000	0.00000	0.00236	0.43799	0.04361	0.00000	0.77496	0.00000	0.00000
Other	0.00000	0.00842	0.00000	0.00000	0.00070	0.01195	0.03233	0.00009	0.00000	0.16823

Figure 2c. Table of Coefficients for Reduced Model								
	Intercept	Sex	AgeAtStart	Height	Systolic	MRW	Smoking	Cholesterol
Cancer	-8.29929	0.30105	0.10394	0.01244	0.00816	-0.00010	0.03223	-0.00193
Cerebral Vascular Disease	-11.52141	0.52788	0.14689	-0.02477	0.02590	0.00150	0.03435	-0.00082
Coronary Heart Disease	-12.98398	1.34471	0.11347	-0.00893	0.02300	0.00681	0.02723	0.00729
Other	-6.21415	0.62672	0.11753	-0.05254	0.01684	-0.00299	0.02989	-0.00174

Figure 2d. Table of p-Values for Reduced Model								
	Intercept	Sex	AgeAtStart	Height	Systolic	MRW	Smoking	Cholesterol
Cancer	0.00000	0.00983	0.00000	0.06934	0.00094	0.96872	0.00000	0.10098
Cerebral Vascular Disease	0.00000	0.00039	0.00000	0.00542	0.00000	0.61608	0.00000	0.55425
Coronary Heart Disease	0.00000	0.00000	0.00000	0.23904	0.00000	0.00898	0.00000	0.00000
Other	0.00000	0.00001	0.00000	0.00000	0.00000	0.34249	0.00000	0.21422

Figure 3a. Table of Coefficients for Full Model										
	Intercept	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol
Cerebral Vascular Disease	3.95056	-0.00849	0.04336	-0.14611	0.02316	0.00381	0.01521	-0.02796	0.00306	0.00114
Coronary Heart Disease	3.1264	0.78984	0.01198	-0.14188	0.02506	0.01452	0.00766	-0.02625	-0.00415	0.00882
Other	12.9869	0.04536	0.01752	-0.24215	0.03797	0.01943	0.00025	-0.05093	-0.00129	0.00002

Figure 3b. Table of p-Values for Full Model										
	Intercept	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol
Cerebral Vascular Disease	0.00000	0.96977	0.00001	0.00000	0.0003	0.65195	0.00057	0.00037	0.62634	0.49123
Coronary Heart Disease	0.00000	0.00009	0.16406	0.00000	0.00001	0.05769	0.06352	0.00019	0.44176	0.00000
Other	0.00000	0.83901	0.06906	0.00000	0.00000	0.02705	0.95797	0.00000	0.83567	0.99091

Figure 3c. Table of Coefficients for Reduced Model								
	Intercept	Sex	AgeAtStart	Height	Weight	Systolic	MRW	Cholesterol
Cerebral Vascular Disease	4.26974	0.02451	0.04161	-0.14855	0.02356	0.01689	-0.02827	0.00118
Coronary Heart Disease	4.58279	0.75036	0.0117	-0.16059	0.02918	0.01369	-0.03021	0.00883
Other	14.97293	0.03095	0.01507	-0.26577	0.0433	0.00829	-0.0561	0.00012

Figure 3d. Table of p-Values for Reduced Model								
	Intercept	Sex	AgeAtStart	Height	Weight	Systolic	MRW	Cholesterol
Cerebral Vascular Disease	0.00000	0.90946	0.00001	0.00000	0.00022	0.00000	0.00028	0.4742
Coronary Heart Disease	0.00000	0.00011	0.15295	0.00000	0.00000	0.00000	0.00001	0.00000
Other	0.00000	0.88521	0.09998	0.00000	0.00000	0.00657	0.00000	0.00657

Figure 4a. Binned Residual Plot of Cancer in Model 1