Brian Holliday

Professor Yuce

Computational Science

22 May 2020

Logistic Regression

Logistic Regression is binary. This means that it is usually used to answer a question that only has two outcomes. This could be yes/no, pass/fail, etc. A logistic regression equation predicts the natural log of the odds of being in one category or another. The logistic function is given by:

$$Y = L/(1 + e^{(a +b*x)})$$

In this equation a, b, and L are constants. Logistic regression produces an s shaped curve. The variety of this S shaped curve depends on the constants. We can get the linear transformation with:

$$\ln((L – y)/y) = a +b*x$$

From this when can y as a piecewise function where y is equation to 1 where A occurs and 0 when B occurs. In this example we will be trying to answer a question related to heart attack risk data. We want to check for the risk factors of heart attacks. Some of these factors age, gender and cholesterol levels.

From the logistic regression we found that men are more likely to have heart attacks than men by about a factor of 2. Age is a risk in having a heart attack as the older you get the highest the chance of having a heart attack. An increase in cholesterol levels is also associated with heart attacks.

Figures:

Figure 1: Frequency Chi-Squared

Comparing Proportions

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Gender by Heart_Attack		
	Heart_Attack(Heart_Attack)		
Gender(Gender)	0	1	Total
F	233 46.60 93.20 52.71	17 3.40 6.80 29.31	250 50.00
M	209 41.80 83.60 47.29	41 8.20 16.40 70.69	250 50.00
Total	442 88.40	58 11.60	500 100.00

Statistics for Table of Gender by Heart_Attack

Statistic	DF	Value	Prob
Chi-Square	1	11.2342	0.0008
Likelihood Ratio Chi-Square	1	11.5397	0.0007
Continuity Adj. Chi-Square	1	10.3175	0.0013
Mantel-Haenszel Chi-Square	1	11.2117	0.0008
Phi Coefficient		0.1499	
Contingency Coefficient		0.1482	
Cramer's V		0.1499	

When we look at our data, we see that the study we have 250 men and 250 women. Men have a high rate of heart attacks with 8.20 percent having had a heart attack. From our chi-squared test we have a p-value of 0.0008, which means we reject the null hypothesis. In this case the null hypothesis is gender is independent of heart attacks. We reject that, so gender is a factor.

Figure 2: Relative Risk

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	2.6887	1.4824	4.8768
Relative Risk (Column 1)	2.4118	1.4089	4.1284
Relative Risk (Column 2)	0.8970	0.8411	0.9566

Sample Size = 500

We can see from the relative risk chart men are 2.68 times as likely to have a heart attack as a woman.

Figure 3: Log Reg Gender

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.6178	0.2512	108.5792	<.0001
Gender	M	1	0.9890	0.3038	10.5994	0.0011

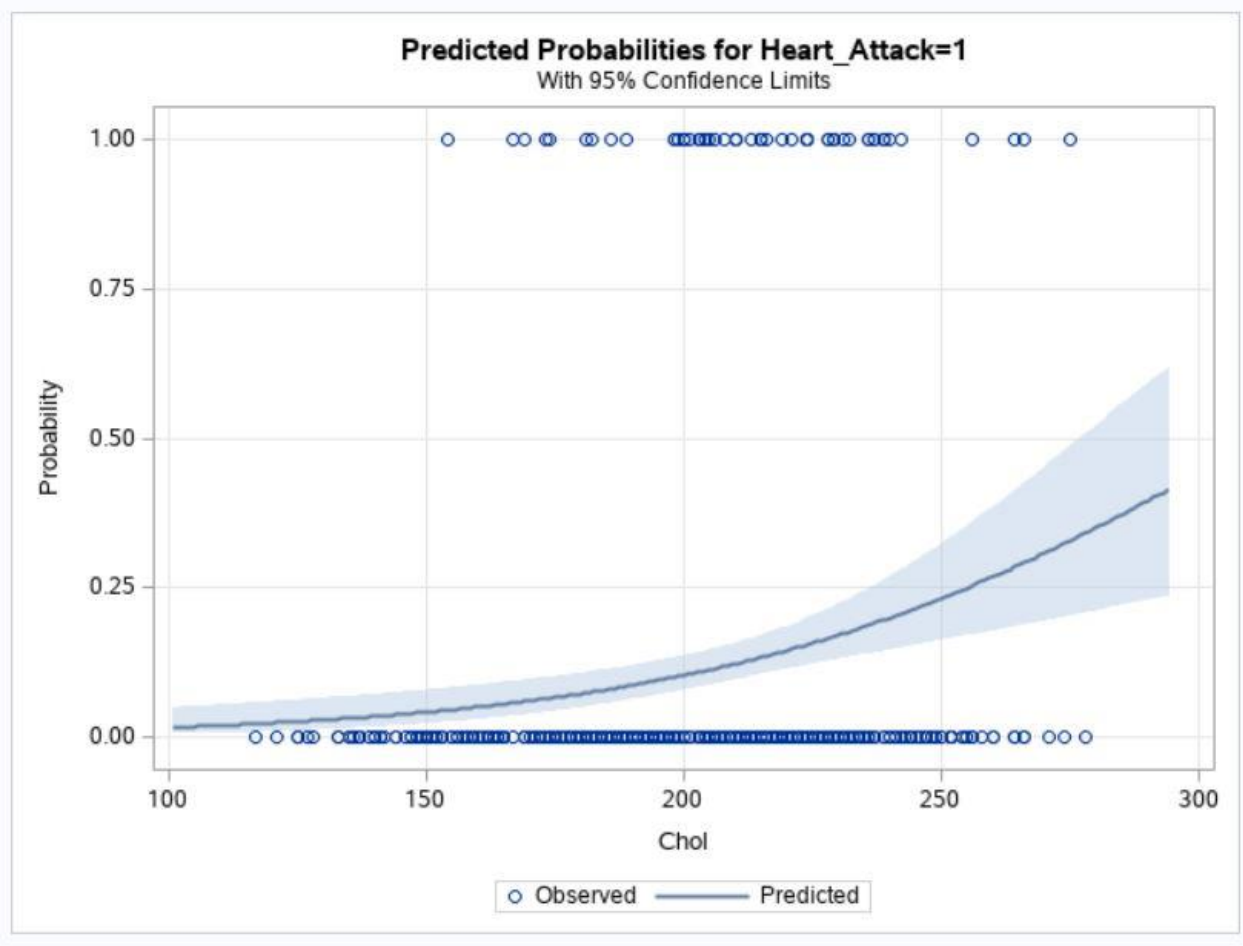
When we run the regression for heart attack equals to gender, we get that the difference in log-odds is e^(0.9880) = 2.689, meaning that that men are 2.689 times more likely to have a heart attack than men.

Figure 4: Log Reg Chol

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.9979	1.0494	32.6651	<.0001
Chol	1	0.0192	0.00488	15.4836	<.0001

In this log regression, since e^(0.0192) = 1.019, we can say that an increase in 1 point of cholesterol will increase the odds of having a heart attack by 1.019 unit.

Figure 5: Heart Attack Probability Graph



The trend in the graph shows that an increase in cholesterol will increase the likelihood of having a heart attack. At about 200 the chances of having a heart attack increases significantly.

Figure 6: Low to High

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.8032	0.2752	103.7386	<.0001
Chol	High	1	1.2356	0.3213	14.7908	0.0001

We can see when we format the data make a cholesterol high after 200 there is a model. We can say that (e^(1.2356) = 3.44), people with high cholesterol has 3.44 times more likey to have a heart attack.

Figure 7: Age Group Model

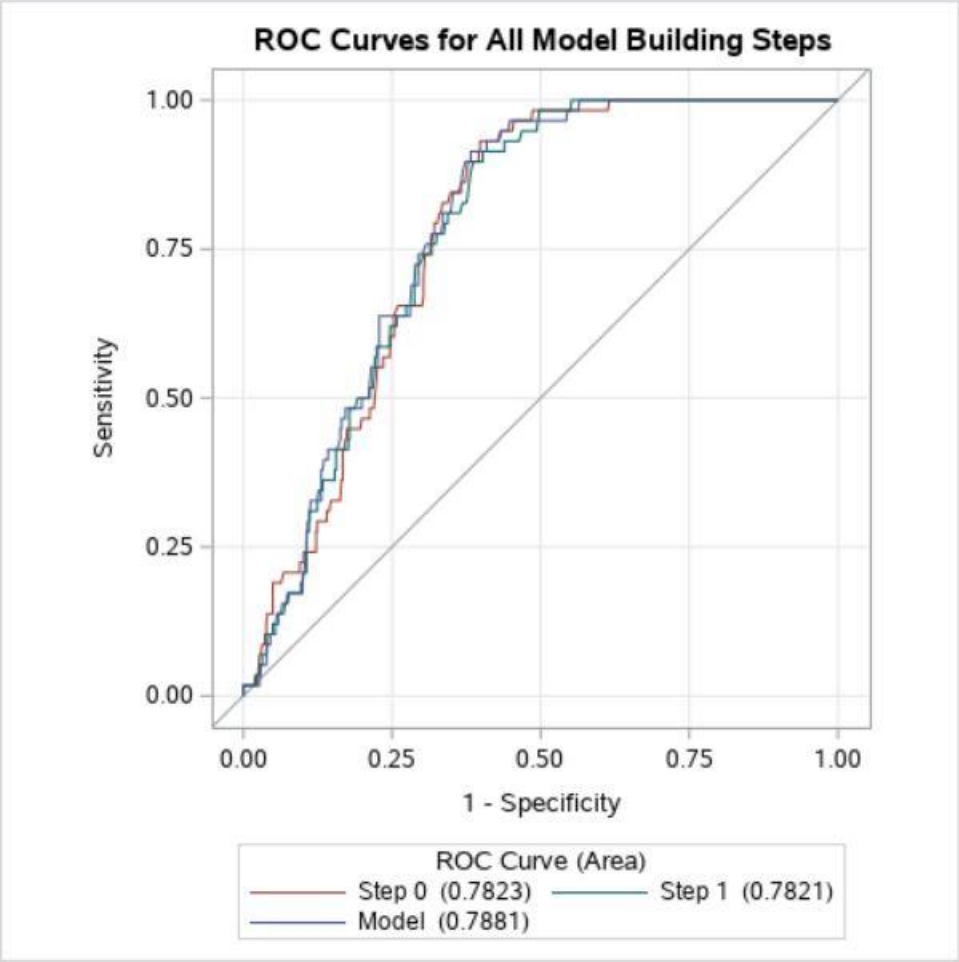
Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.9956	1.2091	43.7300	<.0001
Gender	M	1	1.0028	0.3177	9.9632	0.0016
Age_Group	2:60-70	1	1.4050	0.5148	7.4487	0.0063
Age_Group	3:71+	1	1.9676	0.5092	14.9286	0.0001
Chol		1	0.0193	0.00505	14.5832	0.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	77.6	Somers' D	0.553
Percent Discordant	22.3	Gamma	0.554
Percent Tied	0.1	Tau-a	0.114
Pairs	25636	c	0.776

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Gender M vs F	1.0000	2.726	1.486	5.198
Age_Group 2:60-70 vs 1:< 60	1.0000	4.076	1.601	12.547
Age_Group 3:71+ vs 1:< 60	1.0000	7.154	2.858	21.877
Chol	10.0000	1.213	1.101	1.343

From our model we can see that all the classifications are significant. As for our odds ratios, we can say that men are 2.726 more likely to have a heart attack. The group with the highest risk is the 71+ group that is 7.154 times more likely to have a heart attack than people age lower than 60.

Figure 8: ROC Curve



For the ROC curve we can see