Multiple Logistic Regression on Lung Cancer Data Set

Statistical Methods II Project Report

Scott Williams

1 Introduction

1.1 Data

The data set used to develop a multiple logistic regression model to predict whether a patient has lung cancer was sourced from kaggle. Furthermore, the author on kaggle claims to have collected the data from the online lung cancer prediction system website. The variables present in the data set are gender, age, smoking, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol, coughing, shortness of breath, swallowing difficulty, chest pain, and finally lung cancer. In total, there are 16 variables. We will consider 15 of these to be predictor variables and 1 to be the response variable. In this case, since we are trying to predict the presence of lung cancer, the lung cancer variable will be the response variable and all other variables will be considered as predictor variables. While the names of most of these variables are quite simple to interpret, there are some that are a bit vague. Unfortunately, variables like peer pressure and allergy are not described in detail. In order to get more context for what was asked on the lung cancer survey, we tried to find the online lung cancer prediction system website but it could not be found.

1.2 Goal of the Project

It has long been discussed how technology can aid in the field of medicine. Diagnosing diseases is a perfect subject to apply predictive models to. This kind of method could not only help doctors to have more confidence in their diagnoses but also, as the kaggle author says, provide an effective system for people to know if they are experiencing symptoms of lung cancer at a low cost. We want to build a multiple logistic regression model that can predict whether a patient has lung cancer based on the numerous predictor variables provided with high predictive accuracy.

2 Preparation & Approach

2.1 Programming Environment & Preparation

All the code for this project was done in the SAS OnDemand for Academics Environment. The data set was downloaded from kaggle as a .csv file. Using proc import we loaded our data into SAS. Looking at the data set, we need to do some work to make it usable for logistic regression. For example, the two values of gender in the .csv are M or F and the two values for lung cancer are yes or no. All other predictor variables but age are encoded as 1 for no or 2 for yes. This data set is almost entirely comprised of qualitative data. The only quantitative data that is present is age. Using if then statements in SAS, we create indicator variables for the qualitative data that are coded 0 for no and 1 for yes. Furthermore, male is encoded as 0 and female is encoded as 1. The lung cancer diagnosis follows this same convention.

2.2 Model Building - Selection & Analysis

Since the lung cancer variable has binary outcome, there are different response functions we could choose. Although both the probit mean response function and complementary log-log response function could be used, here we will be using the logistic mean response function. Therefore, our model will be of the form:

$$\pi_i = \frac{exp(X_i'\beta)}{1 + exp(X_i'\beta)}$$

It should be noted, all figures and SAS output described from this point on will be included in the section at the end titled Figures and SAS Output. Since we are trying to build a model to predict whether a patient has lung cancer, the cancer variable is our response variable. The other 15 variables are used as predictor variables. To aid in finding the best model given our data, we compare the results of stepwise selection, forward selection, backward elimination, and best subsets. Using slentry = 0.05 and slstay = 0.1, stepwise selection, forward selection, and backward elimination all return the same model. The nine variables included are allergy, difficulty swallowing, cough, chronic disease, fatigue, peer pressure, smoking, yellow fingers, and alcohol. To further confirm that this is the best model for us to use, we look at the SAS output for the best subsets and use the BEST=1 option to return only the best model for each number of variables included. Since the selection methods return a nine variable model, we look at the best subset output for nine variables which displays the same model as the selection methods. After fitting the model, we obtain these parameter estimates:

1. intercept = -6.6107
2. smoke = 1.4536
3. fingers = 1.7408
4. peer = 1.8742

5. disease = 2.6947
6. fatig = 2.8704
7. aller = 1.8341
8. alc = 1.7514
9. cough = 3.0651
10. swall = 3.4264

Now that we have obtained and fit the best model for our data, let's further examine the adequacy of the model with diagnostic plots.
For logistic regression, we plot the deviance and Pearson residuals against the estimated probability with a lowess smoothing curve on the same plot. These plots should display an approximately horizontal line at zero. If the plot doesn't display this, it could be a sign that the model isn't adequate. In our residual plots, we do obtain an approximately horizontal line at zero, however, there is some fluctuation as the line progresses towards 1.0 on the x axis which could be a bad sign for the adequacy of the model.
We also examine the leverage values, the delta chi-square, and delta deviance statistics to detect outliers in x and influential cases respectively. Examining the plots from SAS, there are quite a few cases that cause spikes meaning we have numerous outliers in x as well as influential cases. In particular, cases 38 and 208 are the most extreme outliers in x and case 30 is the most extreme influential case.
Finally, since we do not have repeated data, we test for lack of fit using the Hosmer-Lemeshow Goodness of Fit Test.

1. Null hypothesis
$$H_0 : E(Y) = \frac{exp(X_i^{'}\beta)}{1 + exp(X_i^{'}\beta)}$$
2. Alternative hypothesis
$$H_a : E(Y) \neq \frac{exp(X_i^{'}\beta)}{1 + exp(X_i^{'}\beta)}$$
3. Test statistic: $G^2 = 2.2522$
4. $\chi^2_{0.05,8} = 15.507$
5. Decision criteria: Reject H_0 if $G^2 > \chi^2_{0.05,8}$

For the Hosmer-Lemeshow test, the number of groups to categorize the data is automatically chosen by SAS. In our case, SAS chose c = 10. For the decision, here we have that $2.2522 > 15.507$ which is obviously false so we fail to reject the null hypothesis. Furthermore, examining p-value = 0.9723 leads us to the same conclusion. Based on the results of the Hosmer-Lemeshow Goodness of Fit

Test, we do not have evidence to say that there is a lack of fit for the model.
Finally, we would like to measure how well our model did at classifying whether a patient had lung cancer. To do this, we calculate the amount of false positives and negatives as well as the amount of true positives and negatives. Using a cutoff point of 0.5, we measure:

1. True Negatives: $28/39 = 0.717$
2. False Negatives: $6/270 = 0.022$
3. True Positives: $264/270 = 0.977$
4. False Positives: $11/39 = 0.282$
5. Altogether, $17/309$ predictions would be incorrect. In other words, we have a prediction error rate of 0.055

These are quite good results. We have great success in classifying true negatives and true positives. However, the two groups of incorrect predictions have quite unbalanced error rates. Normally, it is desirable to see that these error rates are more balanced. It is known that as the cutoff point is shifted further away from the optimal cutoff point in either direction, the error rates are less balanced. Further experimentation would need to be done with the cutoff point to achieve truly optimal results with balanced errors. We can visualize these results with an ROC curve. The area under the ROC curve is always bound between (0, 1) and we want to have a model that results in a larger area under the curve. In our case, we got that the area under the curve = 0.9646 which is a fantastic result. This implies that our model has very good predictive power for classifying whether a patient has lung cancer.
Now that we know we have an appropriate model with high predictive power, we can provide the interpretation of the odds ratio for each of the parameters in the model. We can see from the SAS output that all odds ratios are greater than one so they will all contribute to an increased chance of having lung cancer.

1. When others are constant, compared to nonsmokers, the odds of having lung cancer increase by 327.8%
2. When others are constant, compared to patients without jaundice in the fingers, the odds of having lung cancer increase by 470.2%
3. When others are constant, compared to patients without peer pressure, the odds of having lung cancer increase by 551.6%
4. When others are constant, compared to patients without chronic disease, the odds of having lung cancer increase by 1380.1%
5. When others are constant, compared to patients without fatigue, the odds of having lung cancer increase by 1664.4%

6. When others are constant, compared to patients without allergy, the odds of having lung cancer increase by 526%

7. When others are constant, compared to patients without alcohol use, the odds of having lung cancer increase by 476.3%

8. When others are constant, compared to patients without a cough, the odds of having lung cancer increase by 2043.7%

9. When others are constant, compared to patients without difficulty swallowing, the odds of having lung cancer increase by 2976.7%

Obviously these values are quite extreme, however, they do align with known medical knowledge. For example, at first it might seem surprising that yellowing of the fingers could lead to a 470.2% increased chance of having lung cancer. However, jaundice is known to be a symptom of advanced stage lung cancer so this result is actually reasonable.[1]

3 Conclusion

3.1 Final Thoughts

Although a model with high predictive power for determining whether a patient has lung cancer has been constructed, some of the variables that were left out were surprising. In particular, it was surprising that age was not included in any of the models from the selection processes. Lung cancer tends to be a disease that shows up more in elderly patients. According to the American Cancer Society, the average age of people diagnosed with lung cancer is around 70.[2]

4 Figures and SAS Output

Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	aller		1	1	33.1960		<.0001
2	swall		1	2	27.0768		<.0001
3	cough		1	3	14.9394		0.0001
4	disease		1	4	11.6897		0.0006
5	fatig		1	5	13.6265		0.0002
6	peer		1	6	9.9255		0.0016
7	smoke		1	7	6.8424		0.0089
8	fingers		1	8	6.1623		0.0131
9	alc		1	9	6.6306		0.0100

Figure 1: Results from stepwise selection

Summary of Forward Selection					
Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	aller	1	1	33.1960	<.0001
2	swall	1	2	27.0768	<.0001
3	cough	1	3	14.9394	0.0001
4	disease	1	4	11.6897	0.0006
5	fatig	1	5	13.6265	0.0002
6	peer	1	6	9.9255	0.0016
7	smoke	1	7	6.8424	0.0089
8	fingers	1	8	6.1623	0.0131
9	alc	1	9	6.6306	0.0100

Figure 2: Results from forward selection

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	age	1	14	0.4129	0.5205
2	chest	1	13	0.6487	0.4206
3	gen	1	12	0.3813	0.5369
4	anx	1	11	0.8698	0.3510
5	wheeze	1	10	0.9195	0.3376
6	breath	1	9	1.4802	0.2237

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.6107	1.2677	27.1956	<.0001
smoke	1	1.4536	0.6534	4.9481	0.0261
fingers	1	1.7408	0.6397	7.4067	0.0065
peer	1	1.8742	0.6372	8.6526	0.0033
disease	1	2.6947	0.7618	12.5119	0.0004
fatig	1	2.8704	0.6719	18.2497	<.0001
aller	1	1.8341	0.7237	6.4224	0.0113
alc	1	1.7514	0.7117	6.0560	0.0139
cough	1	3.0651	0.8369	13.4141	0.0002
swall	1	3.4264	0.9797	12.2333	0.0005

Figure 3: Results from backward elimination

Regression Models Selected by Score Criterion		
Number of Variables	Score Chi-Square	Variables Included in Model
1	33.1960	aller
2	57.4945	aller swall
3	74.4516	aller cough swall
4	88.9361	fingers fatig aller alc
5	102.1009	fingers fatig aller alc swall
6	108.0146	fingers fatig aller alc cough swall
7	111.8642	fingers disease fatig aller alc cough swall
8	116.1332	smoke fingers disease fatig aller alc cough swall
9	118.6223	smoke fingers peer disease fatig aller alc cough swall
10	120.2155	smoke fingers anx peer disease fatig aller alc cough swall
11	121.8916	smoke fingers anx peer disease fatig aller wheeze alc cough swall
12	122.7257	smoke fingers anx peer disease fatig aller wheeze alc cough breath swall
13	123.1559	smoke fingers anx peer disease fatig aller wheeze alc cough breath swall chest
14	123.5184	age smoke fingers anx peer disease fatig aller wheeze alc cough breath swall chest
15	123.5190	gen age smoke fingers anx peer disease fatig aller wheeze alc cough breath swall chest

Figure 4: Results from best subsets

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.6107	1.2677	27.1956	<.0001
smoke	1	1.4536	0.6534	4.9481	0.0261
fingers	1	1.7408	0.6397	7.4067	0.0065
peer	1	1.8742	0.6372	8.6526	0.0033
disease	1	2.6947	0.7618	12.5119	0.0004
fatig	1	2.8704	0.6719	18.2497	<.0001
aller	1	1.8341	0.7237	6.4224	0.0113
alc	1	1.7514	0.7117	6.0560	0.0139
cough	1	3.0651	0.8369	13.4141	0.0002
swall	1	3.4264	0.9797	12.2333	0.0005

Figure 5: Model parameter estimates

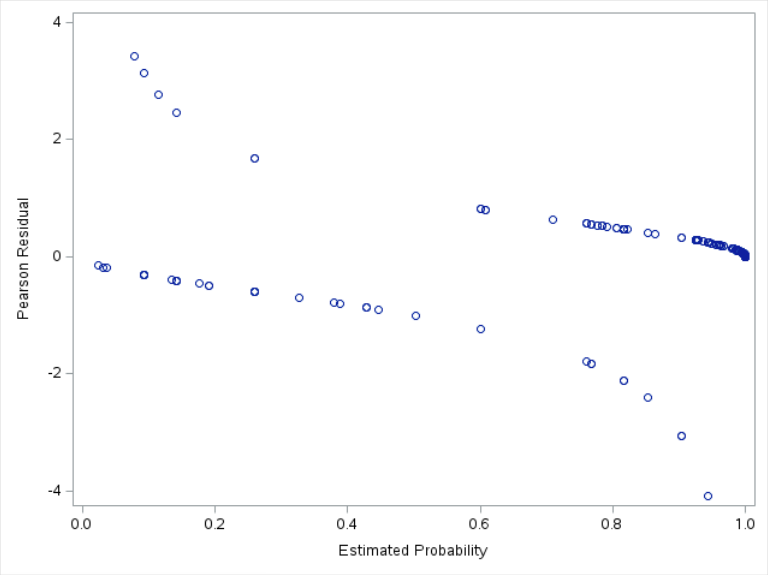


Figure 6: Pearson Residual Plot

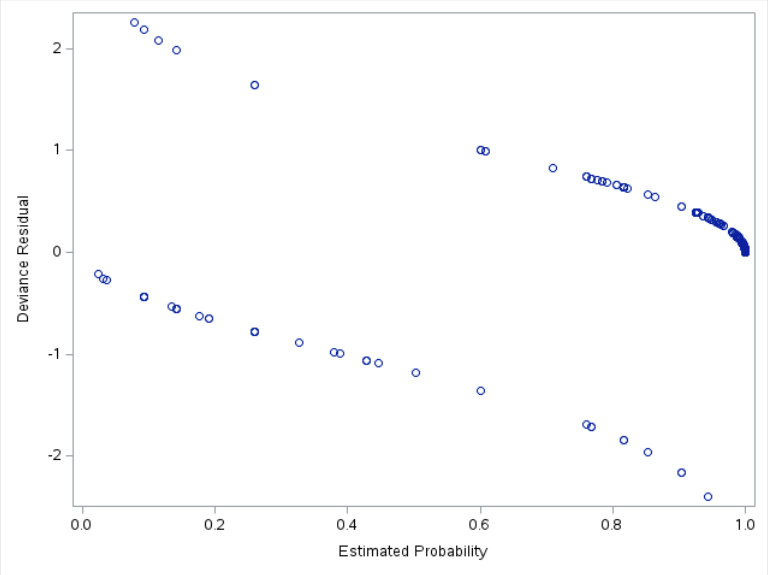


Figure 7: Deviance Residual Plot

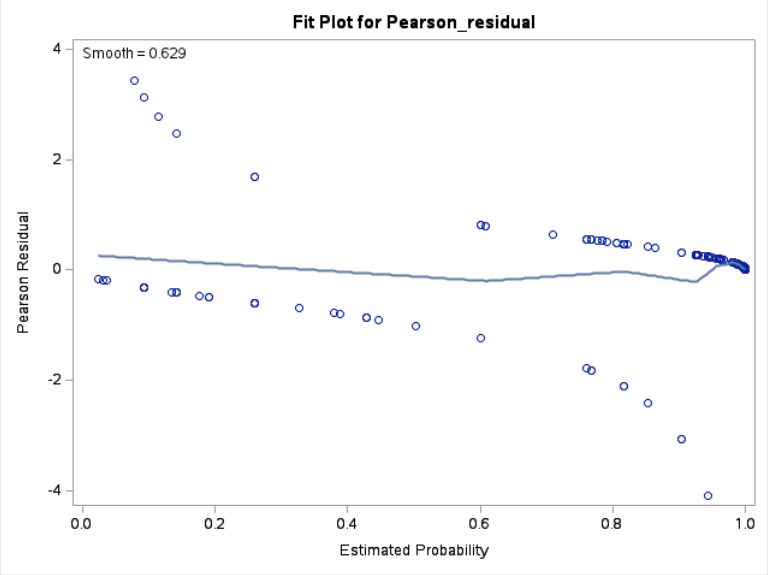


Figure 8: Pearson Residual Plot with LOESS Smooth

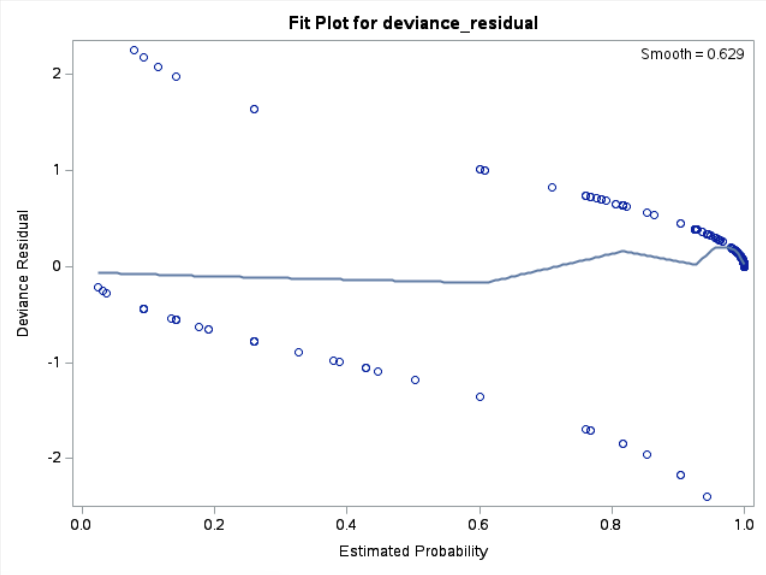


Figure 9: Deviance Residual Plot with LOESS Smooth

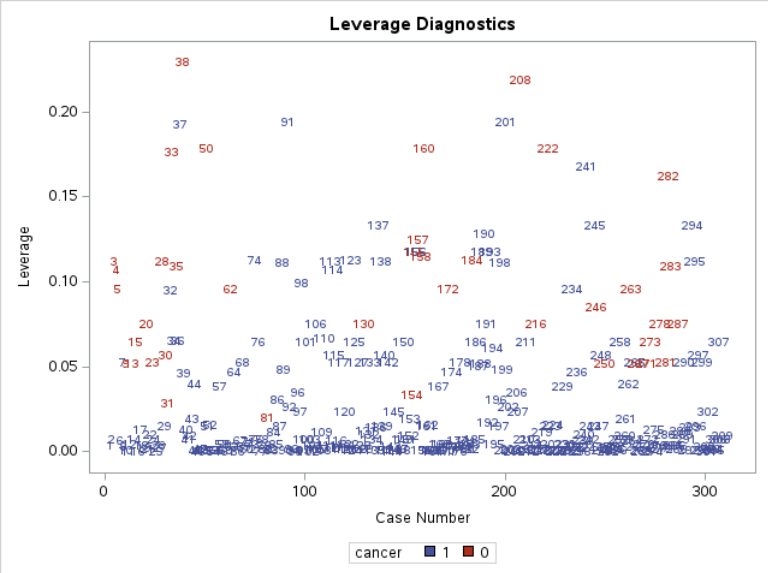


Figure 10: Leverage against Case Number to detect outliers in x

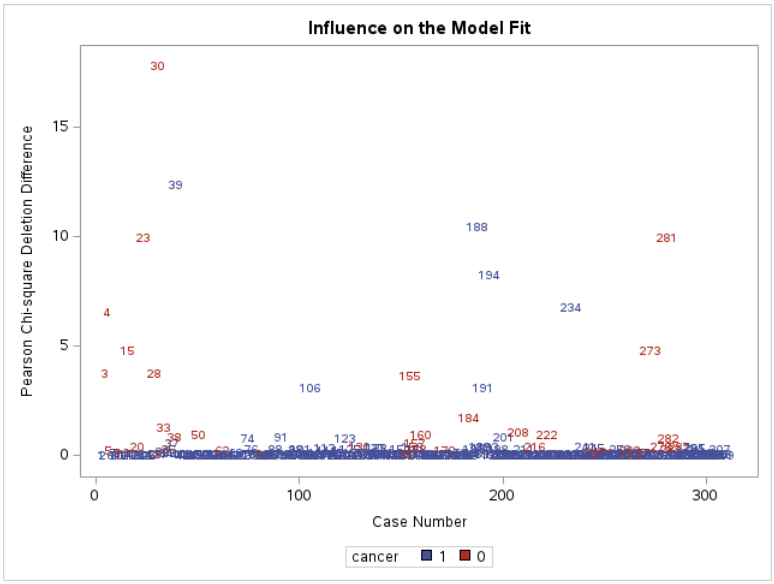


Figure 11: Delta Chi-Square to detect influential cases

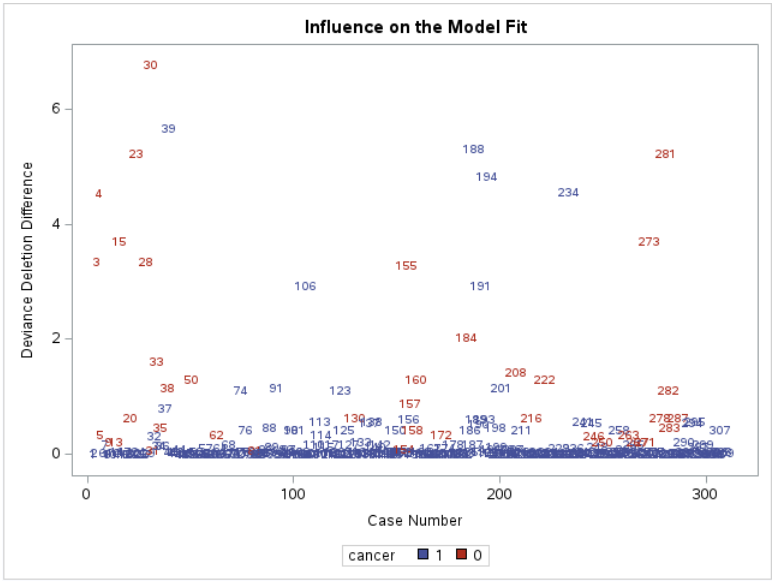


Figure 12: Delta Deviance to detect influential cases

Partition for the Hosmer and Lemeshow Test					
		cancer = 1		cancer = 0	
Group	Total	Observed	Expected	Observed	Expected
1	33	6	6.44	27	26.56
2	34	26	25.26	8	8.74
3	33	29	30.48	4	2.52
4	23	23	22.46	0	0.54
5	31	31	30.62	0	0.38
6	31	31	30.84	0	0.16
7	34	34	33.94	0	0.06
8	33	33	32.98	0	0.02
9	32	32	31.99	0	0.01
10	25	25	25.00	0	0.00

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
2.2522	8	0.9723

Figure 13: Results of the Hosmer-Lemeshow Goodness of Fit Test

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
smoke	4.278	1.189	15.399
fingers	5.702	1.628	19.976
peer	6.516	1.869	22.714
disease	14.801	3.325	65.881
fatig	17.644	4.728	65.845
aller	6.260	1.515	25.858
alc	5.763	1.428	23.251
cough	21.437	4.157	110.542
swall	30.767	4.510	209.878

Figure 14: Odds Ratio Estimates

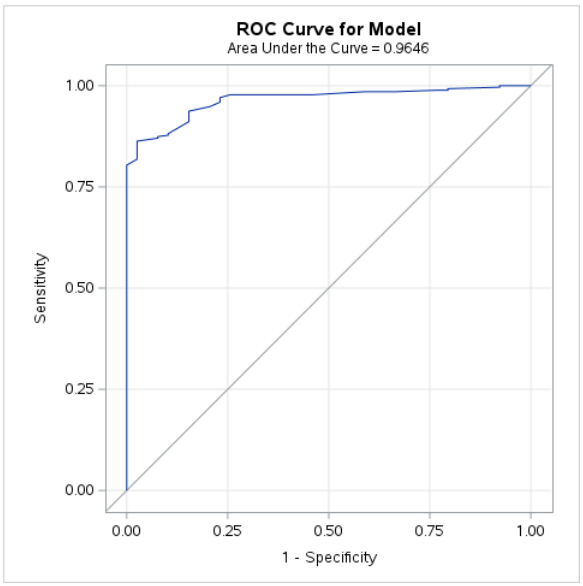


Figure 15: ROC Curve

Number of Observations Read	309
Number of Observations Used	309

Response Profile		
Ordered Value	cancer	Total Frequency
1	1	270
2	0	39

Figure 16: Total observations and amounts of 1's and 0's

cancer=0 predicted=0

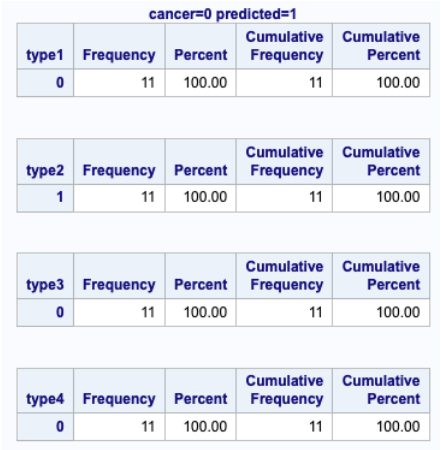
type1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	28	100.00	28	100.00

type2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	28	100.00	28	100.00

type3	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	28	100.00	28	100.00

type4	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	28	100.00	28	100.00

Figure 17: Measuring the amount of true negatives



cancer=0 predicted=1

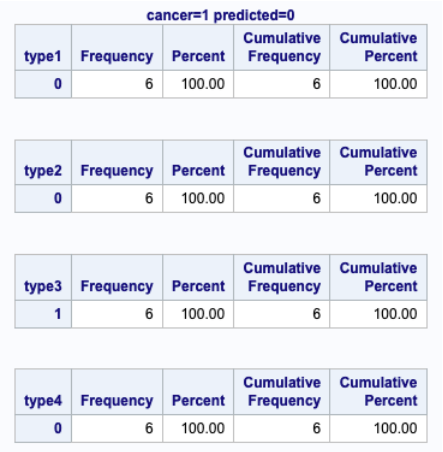
type1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	11	100.00	11	100.00

type2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	11	100.00	11	100.00

type3	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	11	100.00	11	100.00

type4	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	11	100.00	11	100.00

Figure 18: Measuring the amount of false positives



cancer=1 predicted=0

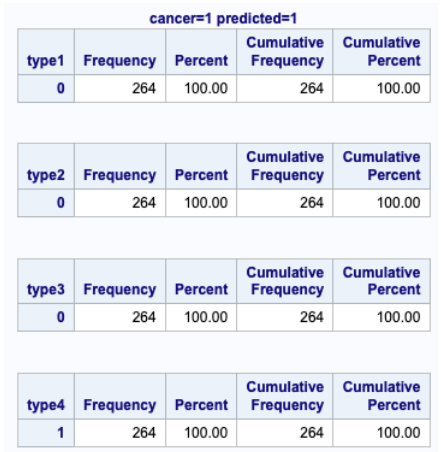
type1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	6	100.00	6	100.00

type2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	6	100.00	6	100.00

type3	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	100.00	6	100.00

type4	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	6	100.00	6	100.00

Figure 19: Measuring the amount of false negatives



cancer=1 predicted=1

type1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	264	100.00	264	100.00

type2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	264	100.00	264	100.00

type3	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	264	100.00	264	100.00

type4	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	264	100.00	264	100.00

Figure 20: Measuring the amount of true positives

5 References

1. Markman M. Lung Cancer. Cancer Treatment Centers of America. Published April 9, 2019. https://www.cancercenter.com/cancer-types/lung-cancer/symptoms

2. American cancer society. Lung Cancer Statistics | How Common is Lung Cancer? www.cancer.org. Published January 12, 2023. https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html