Build a Breast Cancer Binary Logistic Regression Model Using SAS Studio

Didem Bulut Aykurt

MIS500-1 – Foundations of Data Analytics

Colorado State University-Global Campus

Dr. Steve Chung

August 28, 2022

Introduction

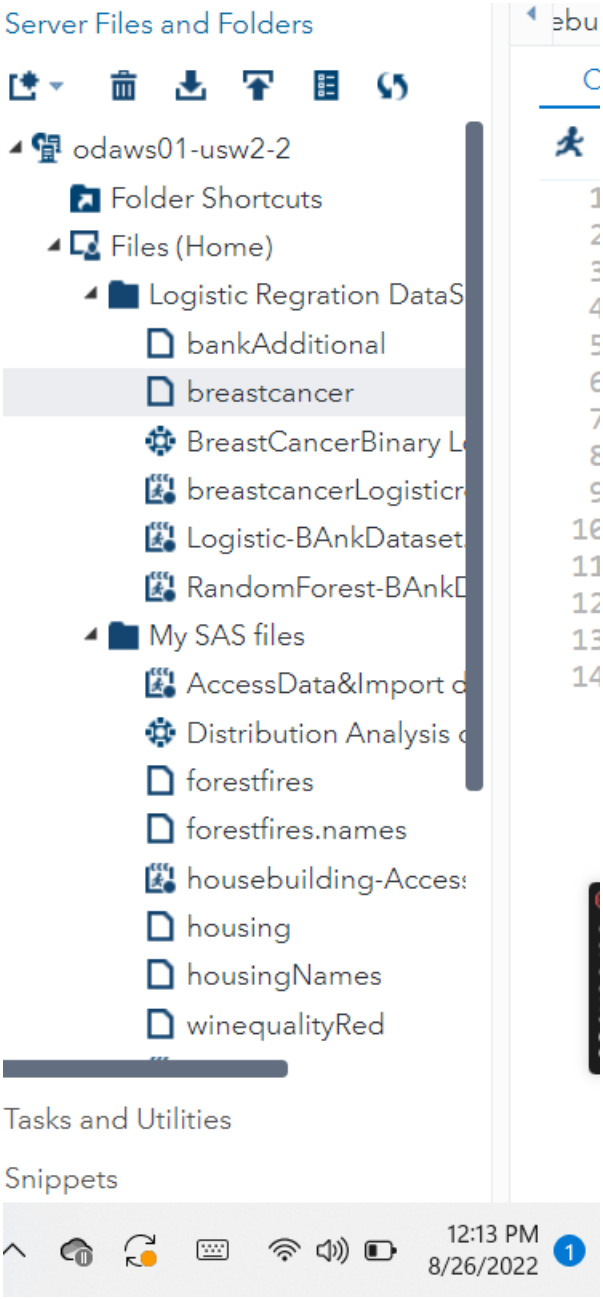
In this case, I will build a binary logistic regression model with The Breast Cancer dataset by a dependent variable "Class" (no-recurrence-event, recurrence-events) with the independent variable "age" in SAS Studio. The data set is publicly available at the UCI Machine Learning repository. I would like to see how the "age" of the independent variable affects breast cancer because breast cancer is common cancer as one in eight women in the USA. The breast cancer data domain from the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia, on 11 July 1988. Additionally, this data set includes nine attributes, the class attribute, and 286 instances.

- Class: no-recurrence-events, recurrence-events
- age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
- menopause: lt40, ge40, premeno.
- tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,
    - 45-49, 50-54, 55-59.
- inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,
    - 27-29, 30-32, 33-35, 36-39.
- node-caps: yes, no.
- deg-malig: 1, 2, 3.
- breast: left, right.
- breast-quad: left-up, left-low, right-up, right-low, central.
- irradiat: yes, no.

Access and import the "breast-cancer" CSV file into SAS studio

The first step to uploading CSV data from my computer into SAS Studio is right click on the existing folder “Logistic Regression DataSet” under the main list “Files(Home)” and uploading a file from my laptop. Now is the “breast-cancer.csv” dataset online on SAS Studio; change the file name right, click then rename “breastcancer” figure 1 has details.

Figure 1
To upload and import the breastcancer data set in SAS Studio.

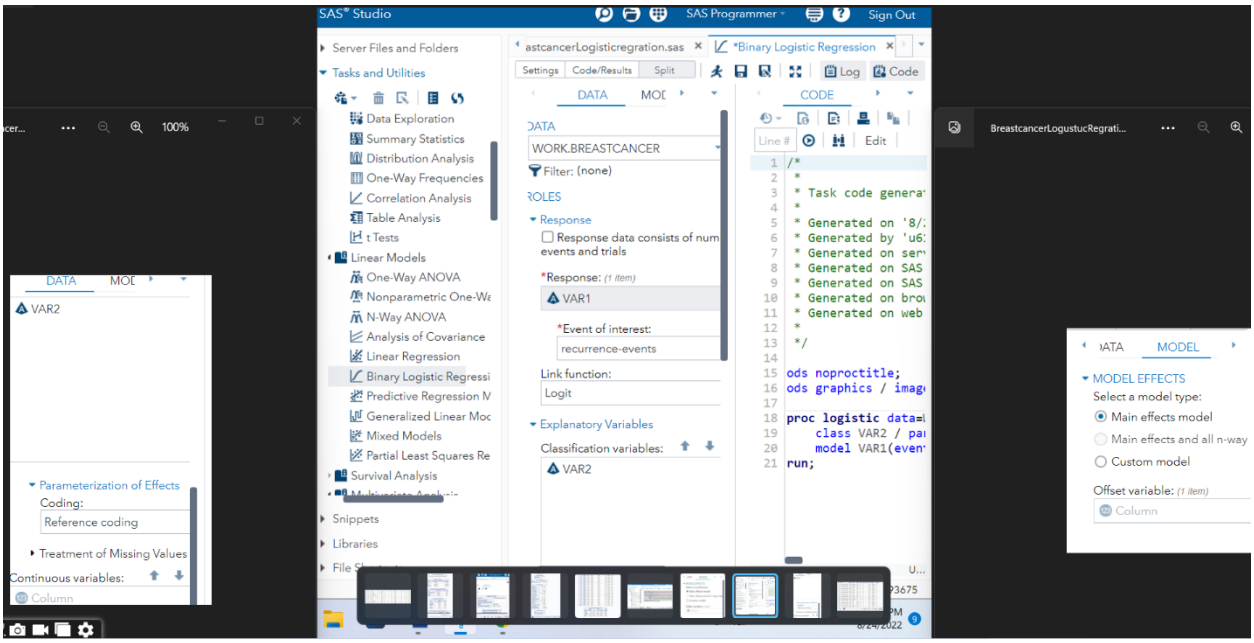


To apply a logistic regression model to the "breastcancer" dataset.

Now the breastcancer dataset is ready to analyze, and I will do the first step to click "Tasks and Utilities" on the main menu and then double click "Binary Logistic Regression" under the "Linear Model." It will pop up a new Logistic regression page. The middle of the page has a "DATA" organizer to choose the response variable VAR1(Class) and the classification variable VAR2(age) and tick "Reference coding" under Parameterization of Effects-Coding. The next step is to edit the "Model" default setting "Main effects model" to a variable of VAR1. Figure 2 shows the detail.

Figure 2

The logistic regression model of breastcancer dataset in SAS Studio.



The result for the logistic regression model of breasthcancer data set in SAS Studio.

8/24/22, 4:23 PM Results: Binary Logistic Regression

Model Information	
Data Set	WORK.BREASTCANCER
Response Variable	VAR1
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	286
Number of Observations Used	286

Response Profile		
Ordered Value	VAR1	Total Frequency
1	no-recurrence-events	201
2	recurrence-events	85

Probability modeled is VAR1='recurrence-events'.

Class Level Information						
Class	Value	Design Variables				
VAR2	20-29	1	0	0	0	0
	30-39	0	1	0	0	0
	40-49	0	0	1	0	0
	50-59	0	0	0	1	0
	60-69	0	0	0	0	1
	70-79	0	0	0	0	0

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	350.048	355.843
SC	353.704	377.779
-2 Log L	348.048	343.843

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.2050	5	0.5203
Score	3.9977	5	0.5497
Wald	3.4822	5	0.6261

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
VAR2	5	3.4822	0.6261

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.6094	1.0954	2.1586	0.1418
VAR2	20-29	1	-11.8288	828.1	0.0002	0.9886

8/24/22, 4:23 PM Results: Binary Logistic Regression

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
VAR2	30-39	1	1.2730	1.1464	1.2329	0.2668
VAR2	40-49	1	0.7621	1.1193	0.4636	0.4959
VAR2	50-59	1	0.5656	1.1199	0.2551	0.6135
VAR2	60-69	1	0.7538	1.1331	0.4426	0.5059

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
VAR2 20-29 vs 70-79	<0.001	<0.001	>999.999
VAR2 30-39 vs 70-79	3.571	0.378	33.782
VAR2 40-49 vs 70-79	2.143	0.239	19.221
VAR2 50-59 vs 70-79	1.761	0.196	15.808
VAR2 60-69 vs 70-79	2.125	0.231	19.580

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	42.8	Somers' D	0.119
Percent Discordant	31.0	Gamma	0.161
Percent Tied	26.2	Tau-a	0.050
Pairs	17085	c	0.559

As a result of the logistic regression model by breastcancer dataset in SAS Studio, the first table shows the data model, and the second, third and fourth tables have data set detail.

- The Model Fit Statistics reports and tests which model fits the data or shows the data fit of the model by AIC (Akaike Corrected) if there are more independent variables and SC for a few separate variable data sets. The -2 Log L can help to compare nested models. The smaller number of results specify better models. Thus, AIC, SC, and -2 Log L don't use here because we applied one model. We don't have another model result to compare those results.
- In the following table of outputs, the likelihood ratio chi-square of 4.2050 with a p-value of 0.5203 shows us that the p-value is more significant than 0.05, which means we can't

reject the null hypothesis. Thus, breast cancer doesn't relate to the age variable, which is the null hypothesis.

- The table Type 3 Analysis of Effects tells the hypothesis test result for each variable. Our model has one variable, VAR2 has a high p-value of 0.6.
- The following table shows the Estimate (coefficients), variables' standard errors, the Wald chi-square, and p-values. Thus, the "age" variable has a few average age limits. One of the ones is (30-39) and has the highest result; the coefficient increases one unit, then the log odds of concession increase by 1.27.
- The Odds Ratio Estimates tell the coefficients as odds ratios are the decrease or increase in odds when the predictor change. If the result is 1 means no difference; if values between 0 to 1 represent a reduction in the probability of the outcome event, and a value more significant than 1 increases the likelihood of the outcome event. Thus, the first age average is (20-29) lower than 1, which decreases the probability, and others that have higher than 1 increase the probability of the outcome.
- The final table contains the result of the relation between predicted probabilities and contemplates the responses; there are four rank results: Somers's'D, Gamma, and Kendall's Tau-a, and c in the table. The Somers'D measure of the dependent and independent variable relationship ranges from values -1 to 1. As our model's result is 0.1 close, the 0 means has no connection between the two variables.