Multivariate Statistical Techniques
CLPS2908, Spring 2019
Prof. Bertram Malle

Homework 4: Logistic Regression

The data set for this homework is available on Canvas. This file contains roughly 200 cases and has six variables. They are, in order:

1. Subject
2. Second suicide attempt (yes = 1; no = 0)
3. Age
4. Sex (1 = male; 2 = female)
5. Depression (depression rating on 0 to 40 scale)
6. SES (socioeconomic status on 1 to 10 scale)

The background to these simulated data is that about 200 people survived a suicide attempt, went through a psychiatric hospitalization, and were tracked over the subsequent 3 years. We want to determine what best predicts whether a person makes a second suicide attempt within these 3 years since release from the hospital.

1. Conduct exploratory data analyses (frequencies, correlations, bivariate plots), report any gross anomalies, and give a synopsis of the relations among the variables. Be sure to include a correlation table of all variables with one another.

2. Run a (parametric) multiple regression, simultaneously entering all predictors. Briefly summarize the results of this analysis (3-4 sentences, no graphs necessary).

3. Observe the predicted values and the residual plots (e.g., raw, standardized, detrended, or whatever you think is useful) to determine whether the parametric regression model provides a good fit to the data.

4. Now run a logistic regression on the same data. First comment on the -2LL numbers that float around in the output:

   (a) What does a -2LL number stand for?

   (b) What is the difference between the -2LL numbers at "Step 0" and at "Step 1"?

   (c) Provide two other indices of the overall success of the regression.

5. Now examine the predictors one by one.

   (a) What contributions does each predictor make to the overall model? (Use plain English as well as appropriate statistics to answer this question.)

   (b) Do we need all predictors in the model?

   (c) Are there any unusual relationships among predictors?

Note: As you know, SPSS doesn't display the (approximated) semi-partial *r*s. You can compute them yourself with the formula I showed in class or you can download the Excel file from the Canvas folder. But note that the Excel file does not take signs into account. You have to transfer the sign from the coefficient B.

6. Examine the classification table of the final model.

(a) What are the two types of error that we can make in prediction? Calculate each error. (You can use the Signal Detection Theory handout in the Canvas folder to get the necessary background.)

(b) One could argue that in the case of suicides we want to minimize cases in which "no suicide" is predicted but the person actually does attempt suicide. We can decrease this error by setting a different threshold from the default 0.5. Pick a reasonable threshold and explore its consequences. What are the benefits and costs of doing that?

7. As usual, write a one-page summary of your analyses, briefly justifying your choice of logistic regression over parametric multiple regression and then focusing on results and interpretation.

The length limit for this homework is 8 pages—but keep the font size reasonable (~Times New Roman 11pt) and the graphs readable.

Extra credit (up to 3 points):

Construct the linear combination u and plot the predicted probability p(event) against the values of u. What is the curve you get?