PRACTICE 6: BINARY OUTCOME

## R software possibilities:

* New method: glm(Y ~ A+B\*X, family=binomial) any formula is valid, where Y is a binary response variable and A and B are factors (qualitative variables) and X a covariate
* Be careful with the default order of factor levels :.
  + Reorder to simplify interpretation: factor(variable, levels=c(nivell1, …, nivellsk))
  + If factor levels are not meaningful include labels for factor levels: factor(variable, levels=c(nivell1, …, nivellsk),labels=c(nom1,…,nomk)).
* step( ) method in R, base on AIC (*Akaike information criteria*) can be used to assess the best model consistent to data.
* Link function by default is logit(probability)=log(probability/(1-probability)).

## CASE 4: US 1992 Presidential Elections

The Elections\_92 archive contains the results of a nationwide sampling of potential voters. Specifically, a sample of 2198 voters nationwide in the US on the attitude towards the presidential election in 1992. The available fields are:

• pres: whether the respondent voted for President in the 1992 election

• age: repondent's age in years

• educ: respondent's years of education

• party: a measure of party preference with seven categories: 0 = strong democrat, 1 = weak democrat, 2 = independent democrat, 3 = independent independent, 4 = independent republican, 5 = weak republican, 6 = strong republican.

• inter: interest in the election: 1 = none, 2 = some, 3 = high.

• close: believes that the election will be close (1 = yes, 0 = no)

• sat: is satisfied with the candidates (1 = yes, 0 = no).

• vote: factor for whether the respondent voted for President in the 1992 election (positive outcome Yes)

• c.age: factor for age (16.29] (29.39] (39.59] (59.91]

• ones: columns including all values as ones.

• c.edu: factor for education (-1,11] (11,12] (12,15] (15,17]

**The response variable is whether or not the individual voted in the 1992 election (1 = yes, 0 = no). You will find an aggregated dataset based on age, called dfage.**

1. Upload the Elections\_92.RData workspace. Take a look at the data before developing the guided exercise in order to fit a good predictive model of the response, to improve readability definition of factor have been incorporated and new variables defined (new variable with ones, indicating the weight of each observation, new factors that contain age and education discretized into groups). Assess by exploratory analysis what are the expected relationships between the response (pres) and the other variables.

2. Estimate Model (M1): Linear predictor with linear AGE covariate.

1. Estimate the model (M1) with disaggregated data (called m1) and aggregated data (called m1a). It is necessary to create a new data frame that collects the information at the level of data added by the classes of the covariate defined by the model M1). You can use the aggregation macro within the scripts available for today's session: carefully monitor the consistency of explanatory variables (covariate) in output data.frame dfage.
2. Interpret the estimated coefficients in terms of logodds, odds, and probabilities.
3. Use the standard car library tools for glm () residual analysis with disaggregated data: residualPlots (m1).
4. Make an overlapping multiple bivariate diagram with the aggregated data format: 1- predicted logodds vs age plus 2-observed logodds vs age. (Age always on abscissas and Observed / predicted values ​​in the scale defined by the link function always in ordinates).
5. Calculate in a new column the empirical / observed logodds: log ((ypos + 0.5) / (m-ypos + 0.5)) = log ((ypos + 0.5) / (yneg + 0.5))
6. Calculate the predicted / adjusted logos in a new column. R already gives you the linear predictor (predict (model)) directly.
7. Make an overlapping multiple bivariate diagram: 1- observed logodds vs Age plus 2- predicted logodds vs Age (Age always on abscissas and Observed / predicted values ​​on the scale defined by the link function always in ordinates).

3. Adjust a logistic regression model with linear predictor including the order 1 and 2 terms of AGE (Models M1-2, M1-3, variants of M1) (nonlinear orders are commonly added with respect to the central tendency of the variable or use the command poly (AGE, k)).

1. Formally contrast the meaning of the 3rd order of age: according to Wald's statistic and the contrast of deviance.
2. Formally contrast the meaning of the 2nd order of age: according to Wald's statistic and the contrast of deviance.
3. Use the command on the disaggregated dataset: marginalModelPlots () in m1 and chosen model.
4. Make a standard diagnosis of the disaggregated model (m2): residualPlots (m2)

4. Consider grouping AGE into categories <30, 30-39, 40-59, 60+. Let it be Model (M2). We will recalculate the m2 model with disaggregated data.

1. Make a standard diagnosis of the disaggregated model (m2c): residualPlots (m2c)
2. Represent the empirical / age-adjusted logits Categorized (handmade diagnostics).

5. Which treatment do you think is most appropriate for the AGE variable: as a covariate (up to which order term) or as a factor. Statistically justify the answer.

6. We will now study the introduction of the variable EDUCATION in the model that already contains the AGE (in its best treatment). To begin with, we will work with EDUCATION as a covariant in the disaggregated dataset format.

1. Add a linear term EDUCATION in the best previous model with AGE.
2. Interpret the estimated coefficient in terms of logodds, odds, and probabilities.
3. Contrast the hypothesis that the EDUCATION effect is linear by introducing a 2nd order term into the model. And is the term order 3 significant? Stop at order 3.
4. Choose the best model used by AGE and the covariate EDUCATION: we will call it (M3). No need to work with the aggregate version unless you want to make an handmade diagnosis.
5. Use the standard residual analysis tools: marginalModelPlots () and residualPlots ().

7. Consider a grouping of years of education (Categorized EDUCATION) into 4 groups: <12, 12, 13-15, 16+. Estimate the logistic regression model with the terms and age-appropriate treatment and the EDUCATION factor. What is the best way to deal with years of education? We will tell him (M4).

1. Try parabolas with/without different parameters for each category of EDUCATION
2. Strictly by inference, what do you think is the most appropriate treatment.

8. What is the best way to deal with years of education, once AGE has been incorporated into the model? Model Result (M6)

9. Add party preferences to the model. Analyze the estimated coefficients and suggest how this variable could be recoded to simplify the interpretation of the model and save a few degrees of freedom (you should see that a maximum of 3 categories is enough). Readjust the model and reinterpret the coefficients of the new coded variable.

10. Enter the variable ‘degree of interest’ in the choices in the model. Contrast the significance of its main effect using: Wald test (if statistical package allows it) and deviance test. Would you say that the main reason young people vote less is that they are not interested in elections?

11. Enter in the model a term for the interaction between ‘degree of interest’ and ‘partisan preferences’. Contrast the significance of the interaction using the deviance test. Explain in detail, if you find evidence, how the interaction works.

12. Is there any evidence that, after introducing the ‘proximity between candidates’ factor into the WORKED TO DATE model (AGE, EDUCATION, PARTY PREFERENCES, DEGREE OF INTEREST), people who believe that elections are adjusted have a higher incidence of voting?

13. Consider satisfaction with nominations. Do people who are not satisfied with the candidacies have a lower incidence of voting? Would the conclusion change if the ‘degree of interest’ in the election was not included in the model?

14. Diagnose your final model.

15. Use your final model to predict a voter’s behavior. Classify as likely voters those individuals with a probability greater than or equal to 0.5. Make a contingency table with the probable vote and the actual vote and analyze it. What is the explicability of the final model?

16. Calculate the pseudo coefficient of determination of the final model and the Naglekerke coefficient. Do a goodness of fit test to the final model. Calculate the goodness of fit using the statistic proposed by Hosmer-Lemershow.

17. Determine the predictive power by analyzing the ROC curve.

18. Consider a modeling methodology that starts with a complete and large model, where the knowledge of the treatment of the variables AGE and EDUCATION is assumed to be known and apply some step () type procedure as a heuristic for choosing the best model. Analyze the results.

19. Write a summary paragraph of the conclusions and what you have learned working with this file.