Langara College

Department of Mathematics and Statistics

DANA 4820

Activity 2

Fall 2023

**Last Name:** \_\_\_\_\_\_Gupta\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**First Name:** \_\_\_\_Rahul\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Student #:** \_100405295\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**\_**

**Question**  – The Students data file at the text website shows responses of a class of social sci- ence graduate students at the University of Florida to a questionnaire that asked about *gender* (1 = female, 0 = male), *age*, *hsgpa* = high school GPA (on a four-point scale), *cogpa* = college GPA, *dhome* = distance (in miles) of the campus from your home town, *dres* = distance (in miles) of the classroom from your current residence, *tv* = average number of hours per week that you watch TV, *sport* = average number of hours per week that you participate in sports or have other physical exercise, *news* = number of times a week you read a newspaper, *aids* = number of people you know who have died from AIDS or who are HIV+, *veg* = whether you are a vegetarian (1 = yes, 0 = no), *affil* = political affiliation (1 = Democrat, 2 = Republican, 3 = Independent), *ideol* = political ideology (1 = very liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = very conser- vative), *relig* = how often you attend religious services (0 = never, 1 = occasionally, 2 = most weeks, 3 = every week), *abor* = opinion about whether abortion should be legal in the first three months of pregnancy (1 = yes, 0 = no), *affirm* = support affirmative action (1 = yes, 0 = no), and *life* = belief in life after death (1 = yes, 2 = no, 3 = undecided).

1. [20] Show all steps of a model-selection (purposeful selection) for choosing a model for predicting *abor*, when the potential explanatory variables are *ideol*, *relig*, *news*, *hsgpa*, and *gender*. (Model 1)

Step 1:

For step 1, any predictor with p-value of less than 0.2 will be considered for the final model. The model with the sole predictor is compared with null model to check the significance, if any.

1. Comparing ideol as sole predictor with null model.

A screenshot of a computer

Description automatically generated

As p-value <0.2, ideol is statistically significant for the main model Model1.

1. Comparing relig as sole predictor with null model

A screenshot of a computer code

Description automatically generated

As p-value <0.2, relig is statistically significant for the main model Model1.

1. Comparing news as sole predictor with null model

A screenshot of a computer

Description automatically generated

As p-value <0.2, news is statistically significant for the main model Model1.

1. Comparing hsgpa as sole predictor with null model

A screenshot of a computer

Description automatically generated

As p-value > 0.2, hsgpa is not statistically significant for the main model Model1.

1. Comparing model with gender as sole predictor with null model

A screenshot of a computer

Description automatically generated

As p-value >0.2, gender is not statistically significant for the main model Model1.

After step 1, the purposeful selection process includes ideol, relig and news as initial explanatory variables.

Initial Main effect model – model6 with ideol, relig and news as predictors. Below is the summary of the model from step 1.

A screenshot of a computer

Description automatically generated

Step 2: perform backward elimination.

From here, p-value will be checked with a value less than 0.05 (5% level of significance)

1. Comparing model 6 – initial model and model 7 – with news removed

A screenshot of a computer code

Description automatically generated

As p-value < 0.05, news will be considered for the final model Model1.

1. Comparing model6 (initial model) and model 8 with relig removed

A screenshot of a computer program

Description automatically generated

As p-value > 0.05, relig will be removed as a predictor and we will keep only ideol and news as predictor variables.

1. Comparing model8 (ideol and news as predictors) and model3 (with ideol removed).

A screenshot of a computer

Description automatically generated

As p-value < 0.05, we will keep model8 with news and ideol as predictors.

Step 3:

Now we will add variables which were not chosen initially.

1. First, we compare model8 with model9 that adds hsgpa condition, which was not one of the initially chosen variables.

A screenshot of a computer

Description automatically generated

As p-value < 0.05, we will keep model9 with news, ideol and hsgpa.

1. Comparing model9 and model99 which has gender added, which was not one of the initially chosen variables.

A screenshot of a computer

Description automatically generated

As p-value >0.05, we will not include gender and will keep model9 with news, ideol and hsgpa as our ideal model for now.

Step 4:

Now, we will check how interactions affect the model.

1. Now model10 adds the interaction between news and ideol and is compared with model9.

A screenshot of a computer code

Description automatically generated

As p-value >0.05, we will not include the interaction term and will keep model9 with news, ideol and hsgpa as it is not significantly better (p-value = 0.951).

1. Comparing model9 and model11which has interaction added between hsgpa and ideol.

A screenshot of a computer code

Description automatically generated

As p-value >0.05, we will not include the interaction term and will keep model9 with news, ideol and hsgpa as it is not significantly better (p-value=0.2563).

1. Now, added model12 that adds interaction between hsgpa and news.

A screenshot of a computer

Description automatically generated

As p-value >0.05, we will not include the interaction term and will keep model9 with news, ideol and hsgpa as it is not significantly better (p-value=0.7938).

The final model Model1 developed for predicting abor has ideol, news and hsgpa explanatory variables as main effects. Below is the summary of Model1.

A screenshot of a computer

Description automatically generated

1. [10] Using an automated tool such as the stepAIC or bestglm functioninR, construct a model to predict *abor*, selecting from the potential explanatory variables in part 1 as explanatory variables.

A screenshot of a computer

Description automatically generated

Based on stepwise backward selection using the stepAIC function, Model1 with three predictors (ideol, news, and hsgpa) is preferred over models with fewer predictors with least AIC values.

1. [10] Using an automated tool bestglm (BIC) function inR, construct a model to predict *abor*, selecting from the potential explanatory variables in part 1 as explanatory variables. (Model 2)

A screenshot of a computer

Description automatically generated

After searching for the subset that has the smallest AIC when used as explanatory variables in a logistic model results the model with ideol and news as predictors.

A screenshot of a computer error

Description automatically generated

The best model selected based on BIC is the one with ideol and news as explanatory variables and has the least AIC values with p-value < 0.05. The above model is now Model2.

1. [10] Create a graph to compare the fitted models for different ideology using Model 2.

A graph of colored lines and numbers

Description automatically generated

Above ideology 1 to ideology 7 are compared in Model2 (news used to predict abor) and it shows that they are parallel and reach at same point after some time. Ideology 1 is at the highest for probability of abor whereas ideology 7 is the least in predicting. The parallelism of curves in horizontal dimension implies that two curves never cross. At each news values, for example, ideology 7 has a lower estimated probability of abor than other news values.

1. [10] Construct an ROC curve, report the area under it for models 1 and 2. What is your conclusion?

A screenshot of a computer code

Description automatically generated

A graph of a model

Description automatically generated

Comparing the ROC curves and the AUC values (area under the curve), Model1 has larger AUC compared to Model2. Based on the AUC values we can use Model1 instead of Model2, but there is not much difference between them (94% for Model1 with 3 predictors and 92% for Model2 with 2 predictors). As there is not much difference, we can say that Model2 is better model with lesser predictors and a AUC value of 0.9223.