Winter 2021 – Stat 141 SL
Esfandiari - Baugh

Problem one. My most recent research interest is examining factors that promote or prohibit academic, social, emotional, and cultural adaptation of STEM and non-STEM majors to UCLA. To that end, I collected a data set of around 800 observations from STEM and non-STEM majors. You can find the data set and codebook in the homework folder on week three.

Create a logistic model for the prediction of satisfaction with academics at UCLA from the students' perceptions of how well their high school prepared them for STEM (Q3.2 -factor A), whether the student is international or domestic (factor B), combined effect of factors A and B, whether the student is transfer or not (factor C), sense of belonging to our campus (factor D), and the combined effect of C and D.

Part one: First you need to take the following steps:

a) Draw a histogram of perception of academic confidence (academic), and cut it to two parts, below the median and above the median. Label above the median "high academic confidence" and below the median "low academic confidence". Calculate the frequencies in the two categories you created.
b) Create a table of frequencies for Q3.2 (my high school prepared me for STEM) and then recode by pooling strongly agree with agree and call it agree and pooling strongly disagree with disagree and call it disagree. Then create the table of frequencies for the recoded Q3.2.
c) Create a table of frequency for international and contingency table of frequencies for the recoded for Q3.2 with international. Label recoded Q3.2 as Q3.2r. Make sure the frequencies look OK.
d) Draw a histogram of sense of belonging to our campus (belonging) and make sure it looks OK.
e) Now run the logistic regression model. Remember that if you simply say A*B it will generate the coefficients for A, B, and AB. You do not need to include all the three terms in the model. In instances that you are only interested in the interaction term and not the main effects, then include A:B in the model.
f) Complete the following table

predictor	Odds ratio	95% CI for Odds ratio	Standard error	P-value
1				
2				
.				
.				
Etc.				

g) Write the null and alternative hypotheses?
h) Draw and interpret the interaction plots within context. Remember that you need to tell R which variables are numbers and which ones are factors. Notice that you will have two interaction plots. When you draw the plot, change the order of Q3.2r so that you get (agree, not sure, disagree). If you do not change the order the plot will not be as clear.
i) Draw the plot of odds ratios and summarize the results within context.

Part two

Using the output of the logistic model…

j) Explain what null deviance shows.
k) Explain what the residual deviance shows.
l) Run an intercept only model using the following command:

```
Model<-glm(academicfactor~1)
> summary(model)
```

m) What do you conclude by comparing residual and null deviance from the intercept only model and your full model?
n) Estimate Pseudo R-squared. Remember that it is wrong to interpret R-squared in logistic regression.
o) Estimate the accuracy of the model by creation of the confusion matrix.

Part three

p) Update the model by deleting the interaction effect between sense of belonging and transfer. (**call this model two**)
q) Conduct ANOVA test to find out if there is any statistically significant difference between the models with and without the interaction term.

Part four

r) Check for outliers and influential points.
s) Ares there any bad leverages that you need to worry about? Yes or no and explain why.
t) Create the marginal model plot and explain what it shows.
u) Check the goodness of fit of the model and comment on whether the logistic model is a good fit for the data.

Problem two

1. Using the diabeticsub data posted in the homework folder on week two, predict the odds of diabetic type II (Diabetes), as a function of hypertension, age, family history of diabetic, and the interaction effect between family history of diabetic and age.

2. Using the library caTools , 65% of the data as testing, and median of predicted scores as threshold, estimate the accuracy of the model.

3. Using mean of predicted scores as threshold, re-calculate the accuracy of the model . Compare the accuracy based on using mean and median as the threshold.

4. Create a ROC curve for estimating the accuracy of the model created for the prediction diabetes. Use 65% of the data as testing.

5. Create a contingency table and estimate the accuracy based on the best cut-off estimated based on the ROC curve.

6. Draw a Roc curve showing true positive and false positive rate and explain what it shows.

7. Perform a five-fold cross validation and explain the findings. Use 70% of the data as testing. Is the accuracy resulting from five-fold cross validation better than the one resulting from the confusion matrix created based on the median and mean cutoff from the ROC Curve?

8. Define sensitivity and specificity and show how you can calculate them for the confusion matrix created in part seven. Discuss whether the model does better with respect to sensitivity or specificity?