Homework 3

Measurement and Evaluation of HCC Systems

How to hand in this homework

- Please email the homework to me as a PDF.
- Late assignments get a penalty of 20% when submitted after the deadline, plus an additional 10% per hour late.
- Make sure you include the R input you used to get to your answer, but do not "dump" the resulting R output on the paper. Copy from the output selectively, and explain it in your own words.
- You may collaborate on this homework, but not copy from others… again, please write your answers in your own words.
- Please include a collaboration statement that says: "I collaborated on this homework with [name]." or "I worked alone on this homework"

Dataset

For questions 1 and 2, you are going to use the same dataset as in Homework 1. Please refer to Homework 1 for a reminder of how the data was collected, etc.

Question 1. Logistic regression

Our recommender system is obviously less useful if the participant already knew all ten recommendations. For this question we are going to investigate these participants.

a. Create a new dichotomous variable "allknown" that is TRUE if the user already knew all ten recommendations, and FALSE if not.

First of all, we expect that music experts may be more likely to already know all recommendations.

b. Run a logistic regression of "allknown" (Y) on music expertise ("expertise", X). Does expertise have a significant effect? What is the p-value?

c. What is the probability of already knowing all recommendations for someone with expertise = 0? How about for someone with expertise = 4?

d. "The odds of already knowing all the recommendations are predicted to be XX% higher for participants with a 1-point higher level of music expertise." Find XX.
e. Give the confidence interval for the odds ratio of music expertise. Does the confidence interval suggest that music expertise is significant? Why (not)?
f. Use the likelihood ratio test to test the significance of this model. Provide both the chi-square value and the p-value.
g. Is the p-value the same as what you found under 1b? Why (not)?
h. Provide the Nagelkerke R^2 for this model.

Let's expand this model…

i. Run a logistic regression of "allknown" (Y) on music expertise and inspectability (list vs graph view).

Interesting! This is not a significant effect, but participants in the list view condition seem to be somewhat less likely to already know all recommendations than participants in the graph view condition! This is strange, because the recommendations in these conditions are calculated by the same algorithm… the only difference is how the recommendations are presented.

j. "Controlling for music expertise, the odds of already knowing all the recommendations are predicted to be XX% higher for participants in the graph view condition than for participants in the list view condition." Find XX.
k. Give the confidence interval for the odds ratio of inspectability.
l. Use the likelihood ratio test to test whether this model is a significant improvement over our first model. Provide both the chi-square value and the p-value.
m. Provide the Nagelkerke R^2 for this model.
n. Report on the two models that you have produced in the same way as Field's Table 8.2.

Question 2. Poisson regression

Instead of looking at whether participants already knew all the recommendations or not, we can also look at the number of recommendations that they *didn't* know already.

a. Create a new variable "dontknow" that counts the number of recommendations the participant didn't know already (hint: this is 10 - known).
b. Create a histogram of the "dontknow" variable with binwidth = 1. Does this variable look normally distributed? What are other problems with this variable if we wanted to conduct a linear regression?

Let's conduct a regular linear regression anyway…

c. Run a linear regression of "dontknow" (Y) on music expertise and inspectability (list vs graph view). Interpret the b-parameters and the p-values of the predictors.
d. How much of the variance in "dontknow" is explained by the predictors?

e. Find the outliers (standardized residuals greater than 1.96 or smaller than –1.96) of this model. Are these problematic?

Now let's conduct a Poisson regression. Given that the linear regression had a very small R^2, we are going to use family=quasipoisson rather than family=poisson. This generally works better when your model has a low R^2.

f. Run a Poisson regression of "dontknow" (Y) on music expertise and inspectability (list vs graph view). Remember: use family=quasipoisson! Interpret the p-values of the predictors.
g. "Controlling for the effect of inspectability condition, participants with a 1-point higher level of music expertise are predicted to have X.X% fewer unknown recommendations." Find X.X.
h. "Controlling for music expertise, participants in the graph view condition are predicted to have XX% fewer unknown recommendations than participants in the list view condition." Find XX.
i. Give the confidence intervals for the odds ratio of music expertise and inspectability.
j. Use the likelihood ratio test to test the significance of this model against the baseline model. Provide both the chi-square value and the p-value.
k. Provide the Nagelkerke R^2 for this model. Is it better than for the linear regression?
l. Find the outliers (standardized residuals greater than 1.96 or smaller than –1.96) of this model. Are these better than for the linear regression?

Dataset

For question 3, you are going to use a new dataset. For this dataset, we classified 308 Facebook users into 6 "privacy awareness" profiles and 6 "privacy behavior" profiles. The awareness profiles are rather straightforward, ranging from "experts" to "novices" in six gradations. The behavior profiles are more interesting:

- **Privacy Maximizers** take the most precautions, including withholding personal information.
- **Selective Sharers** primarily manage custom friend lists to share content selectively.
- **Privacy Balancers** exhibit moderate levels of privacy management behaviors.
- **Self-Censors** use few of the privacy features, but protect their privacy by withholding information.
- **Privacy Minimalists** use only a few common methods, e.g. only sharing with friends by default.
- **Time Savers/Consumers** use privacy strategies to read posts without being bothered by others.

I will send around the manuscript about this data that we have submitted to the International Journal of Human-Computer Studies. The paper talks about the statistical methods we used to

come up with these two classifications; that part is not relevant for this assignment. Our goal in the assignment is to see if there is a significant relation between privacy awareness and behavior.

Question 3. Chi-square test

a. Create a shaded mosaicplot for the dataset. Hint: When you save your plot as PNG or PDF, you can indicate the size. Make it 10”x15” so that the labels don’t overlap.
b. Based on this plot, do you feel like there is a relation between privacy awareness and privacy behavior?
c. Run a chi-square test on the dataset. Make sure to include expected counts, and standardized residuals. Don’t include Fisher’s exact test (the table is too big for this anyway). Is there a significant relation between privacy awareness and privacy behavior?
d. How many expected counts are below 5? Is this a problem?
e. Based on the standardized residuals, what can you say about the privacy behavior of “experts”?
f. Based on the standardized residuals, what can you say about the privacy awareness of “maximizers”? What can you say about “minimalists”?
g. What is the odds ratio of “experts” being “maximizers” rather than “minimalists”, compared to all other levels of privacy awareness?
h. Report on your findings in the same way as slide 24 of the slides for “Categorical data”.