Matthew Isaac

Rates and Proportions – Final Exam

Due: 8 November 2017

1. (Problem 6.3 from textbook)

These data consist of 3 categorical variables (lake, size, food). Food type is the natural response and has 5 categories (Fish, Invertebrate, Reptile, Bird, and Other). Since the response is multi-category nominal, I began by fitting a baseline logits model with food as the response and lake and size and lake\*size as predictors. The interaction term was non-significant, so I removed that and was left with lake (p = 0.0004) and size (p = 0.0009) as two significant variables in my model.

The prediction equations follow the general format:

(for j = 1,2,…,5, k = 1,2,…,4, and l = 1,2)

I have shown the calculations for one such prediction equation below (for lake (k) = George, and size (l) = gt2.3). The rest of the prediction equations would be obtained in a similarly.

Then, we find that:

(Note that )

The results of all possible prediction equations are plotted in the figures below, as generated by SAS.



I have included these plots because they are an excellent representation of the relationship of length and lake on the alligator’s choice of food. Overall, the probability of birds being the main food of choice is always relatively low regardless of the lake or size of the alligator. However, if an alligator does eat birds the most, it is most likely greater than 2.3 meters. Fish seems to be a popular food choice. There is a higher probability of fish being the main food type if the gator is greater than 2.3 meters long, than if it is less than 2.3 meters long, regardless of the lake. However, at lake Hancock, there doesn’t seem to be much of a difference between the probability of the small and large gators eating fish. The smaller alligators seem to always have a higher probability of eating invertebrates, but at Hancock (again) there is only a small difference of the probability between small and large alligators. Lastly, reptiles (similar to birds) seem to be a less popular food choice for alligators than fish or invertebrates. There is almost no difference between the probability of small verses large alligators eating food at a given lake. There is a slight difference from lake to lake, but this likely has to do with the quantity of ‘other’ food available at each lake. In summary, large alligators seem to like fish the best, and small alligators seem to like invertebrates. However, for some reason, at lake Hancock, the size of the alligator seems to play much less of a role in the food choice than at other lakes.

1. (Problem 6.17 from textbook)

These data consist of 4 categorical variables (gender, location, seat-belt, severity). Severity is the natural response variable because we are interested in finding out how gender, location, and whether the individual was wearing a seatbelt are related to the severity of injury. I would be primarily interested in knowing how a seatbelt affects the severity, but gender and location might be interesting as well. It is also important to note that our response variable (severity) has 5 ordinal levels. Because of this, I have used the cumulative odds ratio model.

After fitting an initial model with all interactions, I performed backwards elimination on the effects in the model. This left me with only the main effects in my model – none of the interactions were statistically significant. Gender, location, and seatbelt were all statistically significant (p < 0.0001 for all three variables).

The plots generated by SAS (below) are good summaries about how gender, location, and seatbelt affect the severity of injury suffered by the individuals in accidents.

Lower severity of injury is obviously desirable. Although in some cases the differences were slight, the predicted probability of having an injury with a severity less than or equal to 1 or less than or equal to 2 was higher if the accident happened in an urban location. One possible explanation for this would be that urban speed limits are generally lower than rural speed limits, so a less severe accident is more likely. Likewise, the probability of having an injury with severity less than or equal to 1 (or less than or equal to 2) is higher if the individual was wearing a seatbelt. In other words, wearing a seatbelt is associated with less serious injuries. Lastly, it appears that men are more likely to have an injury with severity less than or equal to 1 (or 2) than women are. This may be because females are generally smaller than males. It also may have something to do with trends of males and females being located at different places in the car.







1. (Problem 9.15 from textbook)

In this problem, these data contain two variables: Gender and Obesity. Each subject has three measurements (obese/not obese) over the course of 5 years. The predictor variable is Gender, and our response variable is Obesity.

I performed a repeated measures analysis for these data, using generalized estimating equations, with three possible covariance structures: compound symmetry, autoregressive, and unstructured.

In all three cases, I found that the predictor variables, Gender, was not significant in predicting obesity (see output below).

| **Score Statistics For Type 3 GEE Analysis (Compound Symmetry)** | | | |
| --- | --- | --- | --- |
| **Source** | **DF** | **Chi-Square** | **Pr > ChiSq** |
| **Gender** | 1 | 0.54 | 0.4617 |

| **Score Statistics For Type 3 GEE Analysis (Autoregressive)** | | | |
| --- | --- | --- | --- |
| **Source** | **DF** | **Chi-Square** | **Pr > ChiSq** |
| **Gender** | 1 | 0.74 | 0.3896 |

| **Score Statistics For Type 3 GEE Analysis (Unstructured)** | | | |
| --- | --- | --- | --- |
| **Source** | **DF** | **Chi-Square** | **Pr > ChiSq** |
| **Gender** | 1 | 0.57 | 0.4520 |

This indicates that there is no association between gender and the child’s obesity over the three years of measurements.

One other thing to note was that, in each case, there was a correlation around 0.5 between repeated observations of obesity for a given subject. In other words, the obesity of a child one year was moderately correlated with his or her obesity in following years.

1. Breast cancer data

These data consist of 5 variables: center, age group, inflammation, appearance, and survived. We are interested in understanding the factors that are associated with a patient living, so I have performed logistic regression on these data, with “survived” as the natural response.

The first thing I did was cut the insignificant factors out of the model using backwards elimination. As a side note, the results of this elimination were surprising to me. I found that center and appearance were statistically significant, and that inflammation and age group were not significant. I would have guessed that age group would be significant (younger patients stronger, able to fight off cancer better), and that inflammation would also be significant (some indicator of the severity of the tumor). I also hoped that center would not be significant. The fact that center is significant implies that there may be some difference between centers that effects whether the patient dies. Now, admittedly, there may be other demographic variables at play here, but if in fact the care or treatment done at different centers is significant in whether the patient survives there should be more investigation in this area. Healthcare professionals may find this quite alarming.

I next wanted to determine if this model was adequately describing these data. In PROC GENMOD, I calculated the deviance. It is 36.6245 on 31 degrees of freedom. P(X231 ≥ 36.6245) = 0.224. Thus, our model does adequately describe these data. However, looking at the ROC curve, I am a bit skeptical about how well it describes these data. The AUC = 0.6014. Its fair but not great by any means.

The odds ratios will help us understand the underlying trends in these data, and how the center and the appearance are related to survival.

| **Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **Center Boston vs Tokyo** | 0.517 | 0.350 | 0.765 |
| **Center Glamor vs Tokyo** | 0.610 | 0.406 | 0.915 |
| **Appearance Benign vs Malign** | 1.675 | 1.209 | 2.320 |

Patients in the Tokyo center are 1/0.517 = 1.934 (nearly two times!) as likely to survive as patients treated in the Boston center, and 1/0.610 = 1.63 times likely to survive as patients treated in the Glamor center. This is quite surprising.

Patients with tumors with a benign appearance are 1.675 times as likely to survive for three years as people with a malignant tumor appearance. This is expected, although I did think that this odds ratio would be larger.

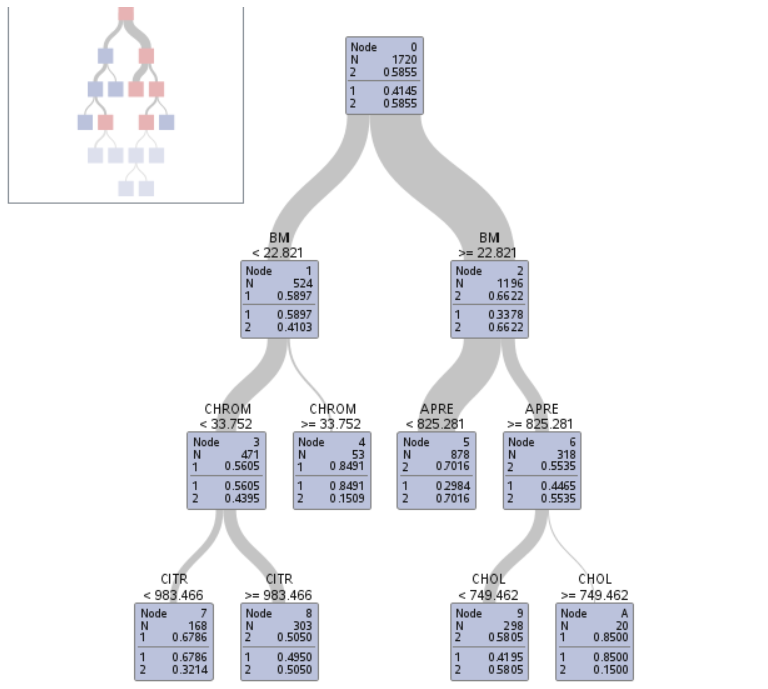
1. Hip Fracture Data

I first fit a classification tree to these data. I didn’t worry too much about any strange values that may be present in the predictor variables as tree-based methods are robust to these outliers. The cross-complexity plot indicates that (following the 1-SE rule) our classification tree should have two leaves. I am hesitant to use a tree that only splits the data once on one variable, but using 9 or 24 nodes results in a more complicated tree and a higher misclassification rate. In order to justify the decision to use a two-node tree, I have compiled the table below which summarizes the accuracy of three possible classification trees.



|  |  |  |  |
| --- | --- | --- | --- |
| Number of Leaves | Cross-Validated PCC | Sensitivity | Specificity |
| 2 | 64% | 0.4334 | 0.7865 |
| 9 | 61% | 0.3562 | 0.7964 |
| 24 | 60% | 0.4348 | 0.7219 |

The 9 leaf tree may still be valuable to the researchers. Although the percent correctly classified is a bit lower, it may provide some insight to the variables which are most useful in predicting hip fractures among the population in question.

Our first split (both on the 2 leaf tree and the 9 leaf tree) is on BMI, with the cutoff being a BMI of 22.821. Individuals with BMI < 22.821 are classified as hip fracture present, while those with BMI ≥ 22.821 are classified as hip fracture absent. If we venture on through the 9 node tree to see which other variables are most useful in classifying hip fractures, we see that CHROM (cutoff of 33.752) and APRE (cutoff of 825.281) are the next two variables used to split the data.

Before performing logistic regression, I investigated the predictor variables to determine which ones might have extreme values or distributions. I noticed one observation that had obscenely high values for a considerable number of the predictor variables. I decided to remove this observation (observation 1555) from the data because the values were so extreme. I then performed log transformations on the variables displayed in the table below.

|  |  |  |
| --- | --- | --- |
| CALCSUPP | ARG | LACT |
| VITDSUPP | ASP | MV\_VITK |
| ADJMMSE | CYS | veg31g |
| PhysicalActivity | GLUTA | veg39g |
| SmokePackperYear | GLY | veg41g |
| AlcoholDrinksperWeek | HIS | veg43g |
| TFA | ISOL | veg44g |
| CHOL | LEUC | veg45g |
| ARE | LYS | veg49g |
| APRE | METH | veg50g |
| VITB2 | PHEN | veg51g |
| B12 | PROL | veg52g |
| BIOT | SER | veg53g |
| EATE | THRE | veg54g |
| EIU | TRYP | veg55g |
| FOLA | TYRO | vkpoor |
| VITK | VAL | allveg |
| *MAG* | ALCO |  |
| ZINC | SORB |  |
| ALA | ACET |  |

The log transformations helped to bring a bring down the extreme variables quite a bit. Performing backwards elimination removed 116 predictor variables.

| **Type 3 Analysis of Effects** | | | |
| --- | --- | --- | --- |
| **Effect** | **DF** | **Wald Chi-Square** | **Pr > ChiSq** |
| **CALCSUPP** | 1 | 5.4917 | 0.0191 |
| **CALC\_SUM** | 1 | 4.9195 | 0.0266 |
| **Weight** | 1 | 45.3763 | <.0001 |
| **Height** | 1 | 37.2095 | <.0001 |
| **GENDER** | 1 | 34.5486 | <.0001 |
| **RAWMMSE** | 1 | 15.0240 | 0.0001 |
| **MOSAC** | 1 | 21.3066 | <.0001 |
| **H2O** | 1 | 4.2692 | 0.0388 |
| **ASH** | 1 | 12.1873 | 0.0005 |
| **CHLOR** | 1 | 8.7988 | 0.0030 |
| **IRON** | 1 | 9.5903 | 0.0020 |
| **POTAS** | 1 | 13.8683 | 0.0002 |
| **CAFF** | 1 | 7.3170 | 0.0068 |
| **EST\_PVUL** | 2 | 9.0789 | 0.0107 |
| **VITK\_SUM** | 1 | 4.4523 | 0.0349 |
| **veg46g** | 1 | 9.7494 | 0.0018 |
| **VITK\_MED** | 1 | 6.9921 | 0.0082 |
| **lSmokePackperYear** | 1 | 6.9471 | 0.0084 |
| **lEATE** | 1 | 12.5290 | 0.0004 |
| **lEIU** | 1 | 8.8269 | 0.0030 |
| **lASP** | 1 | 7.2955 | 0.0069 |
| **lCYS** | 1 | 17.8078 | <.0001 |
| **lLEUC** | 1 | 9.2485 | 0.0024 |
| **lTRYP** | 1 | 18.7482 | <.0001 |
| **lLACT** | 1 | 5.7872 | 0.0161 |
| **lveg39g** | 1 | 10.6064 | 0.0011 |
| **lveg53g** | 1 | 6.9483 | 0.0084 |

The ROC curve for this model gives us an AUC of 0.7335. This indicates that our model is a fair fit – not spectacular, but fair. At a probability level of 0.5, the model gives 67.2% correctly classified. The sensitivity is quite disappointing (49.1), but the specificity is much higher (80.1). Unfortunately, I would imagine that in this specific application the researchers would be most interested in having a high sensitivity – that is correctly predicting the presence of a hip fracture. This would help them to (hopefully) be able to identify and help those most at risk before they suffered a hip fracture. In this case, the researchers might consider using a probability level more around 0.3, which would give a sensitivity of 81.6% (presences correctly classified).

There was a whole slew of odds ratios generated from this model. I have chosen to discuss some of the larger and more notable odds ratios. First, Gender seems to play a significant role. Females are 2.973 times as likely to suffer from a hip fracture as males. Patients with EST\_PVUL = +/+ are 0.673 times as likely to have a hip fracture as patients with EST\_PVUL = -/-. Likewise patients with EST\_PVUL = +/- were 0.670 times as likely to have a hip fracture as patients with EST\_PVUL = -/-. Height was also a surprising odds ratio. For every unit increase in height (meter?), patients were 239.863 times more likely to have a hip fracture. For every unit increase (on the log scale) of EIU, patients were 6.215 times more likely to have a hip fracture, and for every unit increase (on the log scale) of ASP, patients were 27.921 times more likely to have a hip fracture. Lastly, for every unit increase(on the log scale) of LEUC, patients were 0.021 times as likely to have a hip fracture (a decrease in likelihood).

| **Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **Height** | 239.863 | 41.235 | >999.999 |
| **GENDER F vs M** | 2.973 | 2.067 | 4.275 |
| **EST\_PVUL +/+ vs -/-** | 0.673 | 0.496 | 0.911 |
| **EST\_PVUL +/- vs -/-** | 0.670 | 0.509 | 0.882 |
| **lASP** | 27.921 | 2.493 | 312.720 |
| **lLEUC** | 0.021 | 0.002 | 0.252 |

Ultimately, the logistic regression model provided us with a slightly higher cross-validated percent correctly classified (67%) than the classification tree (64%). However, the classification tree was comparable in accuracy, and could be selected to classify these data if easy interpretation of the model is needed or desired.