MATH 536 Homework 6

Gustavo Esparza

11/13/2019

1 Vehicle Safety

A Fit a baseline category logit model for this data using no/little importance as the baseline for the response. Interpret the parameters in the logit equation for important vs no/little importance.

Here is a summary of our logit model, using the given baseline. In this instance, coeffecients with index **1** and **2** are related to Important and Very Important, respectively.

```
##                   Estimate Std. Error   z value     Pr(>|z|)
## (Intercept):1  0.6087802  0.3652683  1.666666 9.558083e-02
## (Intercept):2  1.0646582  0.3496452  3.044967 2.327059e-03
## age18-23:1    -1.5877072  0.4028993 -3.940705 8.124246e-05
## age18-23:2    -2.9167514  0.4229223 -6.896660 5.323933e-12
## age24-40:1    -0.4594414  0.4226842 -1.086961 2.770539e-01
## age24-40:2    -1.4386445  0.4158219 -3.459762 5.406540e-04
## sexwomen:1     0.3881281  0.3005117  1.291557 1.965104e-01
## sexwomen:2     0.8130175  0.3210362  2.532479 1.132591e-02
```

Focusing on important vs little/no importance, we have the following log odds equation:

$$log(\pi_I/\pi_{NLI}) = .6087 - 1.5877(Age_{18-23}) - .459(Age_{24-40}) + .388(Sex_F)$$

The intercept of .6087 represents the odds for the baseline category (Men aged greater than 40). Specifically the odds for male drivers greater than 40 ranking important over no/little importance is 1.838 times the respective odds for males less than 40.

We can observe that both age slopes are negative, and the slope for the age group 18-23 is larger in magnitude. This implies that the odds of a driver ranking "important" is smaller than the odds of a driver ranking "no/Little importance" when the driver is younger. Specifically, the odds will be 0.204 and 0.632 times the odds for male Drivers older than 40, for ages 18-23 and 24-40, respectively.

We can also observe that the slope for gender (Female) is .388. This implies that the odds of ranking "important" is greater than the odds of ranking "no/little importance" when the driver is a Female greater than 40 opposed to Males greater than 40. Specifically, the odds for female drivers ranking important over no/little importance is 1.474 the respective odds for males.

i Estimate the probability of no/little importance for women aged 18-23.

The probability of no/little importance for women ages 18-23 is defined as

$$\frac{1}{1 + exp(.608 - 1.588 + .388) + exp(1.06 - 2.916 + .813)}$$

Thus, the probability of no/little importance for women ages 18-23 is 0.5242007.

ii Estimate the probability of no/little importance for men age greater than 40.

The probability of no/little importance for men age greater than 40 is defined as

$$\frac{1}{1+exp(.608)+exp(1.06)}$$

Thus, the probability of no/little importance for men age greater than 40 is 0.1742757.

B Fit a proportional odds models. Interpret the parameters in one of your equations.

Here is the summary for our proportional odds model:

```
##                                      Value Std. Error   t value
## age1>40                         2.23245971  0.2914908  7.658766
## age124-40                       1.14709969  0.2776161  4.131964
## sex1men                        -0.57622616  0.2261865 -2.547571
## No.little.importance|Important  0.04353714  0.2322724  0.187440
## Important|Very.important        1.65497448  0.2556182  6.474401
```

By inspection of the parameters for No/Little Impotance VS Importance, we have the following log proportional odds equation (with signs changed):

$$log(\pi_N/(\pi_I+\pi_V)) = .0435 - 1.147(age_{24-40}) - 2.232(age_{>40}) + .576(Sex_M)$$

Here, the intercept is .0435. This implies that the odds of the baseline category (age 18-23, sex = F) having No/little importance are greater than the odds of the baseline having either Importance or Very Important ranking. Specifically odds of the baseline category having No/little importance are 1.044 times the odds of the baseline having either Importance or Very Important.

Our age slopes are both negative values greater than one, which implies that the odds of choosing no/little importance is smaller for these two age groups when compared to the odds of choosing important/very important. Specifically the odds of choosing no/little importance is 0.3198 and 0.108 times the odds of females age 18-23 choosing the other rankings, for females ages 24-40 and >40, respectively.

The slope related to Sex(Male) is .576, implying that the odds of choosing "No/little importance" is greater for males when compared to the odds of selecting "important/very important". Specifically the odds of males age 18-23 choosing no/little importance over the other rankings is 1.779 times the respective odds for females age 18-23.

i Find the probability of each category for females aged 18-23.

The cummulative probability for each category given the criteria (females aged 18-23) is defined as

$$p(y \leq i) = \frac{e^{\beta_0 i}}{1+e^{\beta_{0i}}}, i = 1, 2$$

Where the cummulative probability for "Very Important" is simply 1. Then, using the difference between these cummulative probabilties allows us to compute the exact probabilties for each category.

Thus the probability of our criteria driver being in category "Very Important" is defined as $1 - p(Y \leq Important) = 0.1604378$.

The probability of our criteria driver being in category “Important” is defined as $p(Y \leq Important) - p(Y \leq NotImportant) = 0.3286797$.

Finally, the probability of our criteria driver being in category “Not Important” is defined as $p(Y \leq NotImportant) = 0.5108826$

```
## No.little.importance            Important       Very.important
##            0.5108826            0.3286797            0.1604378
```

2 Car Accidents

A Considering a male wearing a seatbelt with an accident in an urban location as the baseline. Fit a proportional odds model. State the model and the distribution assumptions.

We are assuming that the model has a multinomial distribution with 5 responses denoted by y_i. We also assume that the slopes remain constant for any accident severity.

Here is the summary for our proportional odds model:

```
##
## Re-fitting to get Hessian
```

```
##                   Value Std. Error   t value
## genderFemale  0.5449006 0.02722450  20.01508
## beltNo        0.8240650 0.02761807  29.83790
## locationRural 0.7732150 0.02693557  28.70609
## one|two       3.3455339 0.03115399 107.38702
## two|three     3.5198969 0.03168379 111.09457
## three|four    5.3869354 0.04419778 121.88248
## four|five     7.2937253 0.09007363  80.97514
```

The model is defined as

$$log((\pi_1 + \cdots + \pi_i)/(\pi_{i+1} + \cdots + \pi_5)) = \alpha_i - .544(Female) - .824(NoBelt) - .773(Rural), i - 1, \ldots, 4$$

Where α_i represents the baseline cummulative probabilities for categories 1-4.

Interpretation of Slopes

Gender: The odds of a Female being in group i or below using a seatbelt in an urban area are 1.7228846 times the respective odds for a male using a seatbelt in an urban area.

belt: The odds of a Male being in group i or below not using a seatbelt in an urban area are 2.2796 times the respective odds for a male using a seatbelt in an urban area.

location: The odds of a Male being in group i or below using a seatbelt in a rural area are 2.1662553 times the respective odds for a male using a seatbelt in an urban area.

B Given the baseline, find the estimated probabilities for each of the response categories.

The cummulative probability for each category given the criteria (males wearing seatbelts in urban areas) is defined as

$$p(y \leq i) = \frac{e^{\beta_0 i}}{1 + e^{\beta_{0i}}}, i = 1, 2, 3, 4$$

Where the cummulative probability for "5-injured and not survived" is simply 1. Then, using the difference between these cummulative probabilties allows us to compute the exact probabilties for each category.

Thus the probability of our baseline subject being in category "1" is defined as 0.9659583.

The probability of our baseline subject being in category "2" is defined as 0.0052903.

The probability of our baseline subject being in category "3" is defined as 0.0241962.

The probability of our baseline subject being in category "4" is defined as 0.0038758.

The probability of our baseline subject being in category "5" is defined as .00067.

Using the **predict** function, here are the estimated probabilties for each response given the input of Males wearing seatbelts in an Urban area:

```
##          one          two        three         four         five
## 0.9659582838 0.0052903414 0.0241962436 0.0038758021 0.0006793291
```

We can see that the predict function and the manual computations are the same.

C Find the cumulative odds ratio of gender, given location and seatbelt use. Interpret this odds ratio.

The odds ratio for gender in this scenario is defined as

$$\frac{odds(Y = i|female)}{odds(Y = i|male)}$$

The model coeffecients give -0.5449006 as the coeffecient relating to gender. Taking the exponent of this value gives us 0.5798995 as the cumulative odds ratio of gender. Thus, for any constant seatbelt and location status, the odds of a female being in or below response group i are 0.5798995 times the equivalent odds for males.

D Find the cumulative odds ratio of seatbelt use, given that the accident occurred in a rural location. Does this differ from accidents in urban locations? If so, why?

The odds ratio for seatbelt use in this scenario is defined as

$$\frac{odds(Y = i|\textbf{No seatbelt and rural})}{odds(Y = i|\textbf{Seatbelt and rural})}$$

The model coeffecients give -0.824065 as the coeffecient relating to no seatbelt and rural location. Although rural status does have an associated coefficient, it will be removed in the odds ratio as it is present for both odds of no seatbelt vs seatbelt. Taking the exponent of this value gives us 0.4386449 as the cumulative odds ratio of seatbelt use in rural areas. Thus, for any constant gender, the odds of a rural non-seatbelt driver being in or below response group i are 0.4386449 times the equivalent odds for seatbelt rural drivers.

Since the location status is being held constant, we would not see a difference in odds for rural and urban drivers because the odds ratio will cancel out both values and leave the seatbelt coefficient remaining.

3 Estimate a Poisson regression model to predict the number of satellites as a function of the given explanatory variables. What should your final model be? Interpret the parameters in your final model. Give a summary of the process you used to determine this model.

Here is the summary for the model explaining the number of satellites as a function of Width:

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -3.3047572 0.54224155 -6.094622 1.096964e-09
## W            0.1640451 0.01996535  8.216491 2.095450e-16
```

We can see that our intercept and predictor appear to be statistically significant within the model, but we should still check for dispersion in our Poisson model. We can do this by comparing the ratio between our residual deviance and the degrees of freedom for the residual deviance. In this case we have a ratio of 3.3209273. As this value is much greater than one, we may be able to fix our issue with overdispersion by adding another predictor to the model that can help explain the variance amongst the data.

Before proceeding with another predictor, we can also test our model for goodness of fit using residual deviance. Our test is defined as follows:

H_0: Current model VS H_A: Saturated Model
Test statistic: χ^2 with 171 degrees of freedom: 567.8785725
P-value: ≈ 0

Thus we proceed to reject our null hypothesis and conclude that our current model does not fit well.

Thus, we we will add the Color as a predictor to see if we can resolve our issue of overdispersion. Here is the summary for this additive model:

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -2.5199832 0.61062880 -4.126866 3.677405e-05
## W            0.1495725 0.02067865  7.233185 4.717944e-13
## C           -0.1694036 0.06184214 -2.739291 6.157188e-03
```

Once again, we can see the predictors appear to be statistically significant. We will once again check for overdispersion with our useful ratio, in this case we have a dispersion ratio of 3.2953019. It appears that adding another predictor did not resolve the issue of overdispersion.

Once again, we can also test our model for goodness of fit using residual deviance. Our test is defined as follows:

H_0: Current model VS H_A: Saturated Model
Test statistic: χ^2 with 170 degrees of freedom: 560.2013236
P-value: ≈ 0

Thus we proceed to reject our null hypothesis and conclude that our current model does not fit well.

In chapter 3 of the text, there is a discussion of grouping data to construct Poisson models. In particular, we have grouped the original data by subsets of width(represented here by the median width) and counted the number of craps in each group, along with the total amount of satellites within the group.The text also presented the means and variances for the satellites found within the group, which showed that it was not constant. To this end, we will include an offset variable defined as the log of the number of cases for each grop; this will hopefully scale the average number of satellites found in each group and therefore assist with preventing overdispersion.

```
##    width cases  Sa
## 1 22.125    14  14
## 2 23.750    14  20
## 3 24.750    28  67
## 4 25.750    39 105
## 5 26.750    22  63
## 6 27.750    24  93
## 7 28.750    18  71
## 8 31.400    14  72
```

Here are the results from our grouped rate regression model:

```
##
## Call:
## glm(formula = Sa ~ width + offset(log(cases)), family = poisson(link = log),
##     data = Crab2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.57339  -1.12316  -0.08125   0.66430   1.24630
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.86962    0.49254  -5.826 5.67e-09 ***
## width        0.14746    0.01806   8.164 3.23e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 72.3772  on 7  degrees of freedom
## Residual deviance:  7.8341  on 6  degrees of freedom
## AIC: 58.279
##
## Number of Fisher Scoring iterations: 4
```

We can once again see that our predictors are statistically significant and our dispersion ratio is now 1.3056846. This ratio iimplies that the issue of overdispersion has been resolved.

We can also test our model for goodness of fit using residual deviance. Our test is defined as follows:

H_0: Current model VS H_A: Saturated Model
Test statistic: χ^2 with 6 degrees of freedom: 7.8341
P-value: 0.2505108

Thus we proceed to fail to reject our null hypothesis and conclude that our current model accurately fits the data.

3.20

A For each age, compute the sample coronary death rates per 1000 person-years, for nonsmokers and smokers. To compare them, take their ratio and describe its dependence on age.

Here is our original table with a new column representing the sample death rate per 1000 person-years:

```
##    deaths personyears smoke   age       rate
## 1       2       18793     0 35-44  0.1064226
## 2      12       10673     0 45-54  1.1243324
## 3      28        5710     0 55-64  4.9036778
## 4      28        2585     0 65-74 10.8317215
## 5      31        1462     0 75-84 21.2038304
## 6      32       52407     1 35-44  0.6106055
## 7     104       43248     1 45-54  2.4047355
## 8     206       28612     1 55-64  7.1997763
## 9     186       12663     1 65-74 14.6884624
## 10    102        5317     1 75-84 19.1837502
```

It is clear that the death rate increases with age for both smokers and non-smokers.

Furthermore, here are the ratio of deaths for nonsmokers to smokers in each age group:

```
## [1] 0.1742903 0.4675493 0.6810875 0.7374306 1.1053016
```

We can observe that smokers have a higher rate of death for younger ages, but the ratio becomes approximately one for older ages.

B Specify a main-effects Poisson model for the log rates having four parameters for age and one for smoking. Explain why this model assumes a constant ratio of nonsmokers' to smokers' coronary death rates over levels of age. Based on (a), would you expect this model to be appropriate?

Here is the summary for our main-effects Poisson model:

```
##               Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) -7.9193658  0.1917601 -41.298296 0.000000e+00
## age45-54     1.4840386  0.1951029   7.606440 2.817498e-14
## age55-64     2.6275364  0.1837268  14.301322 2.146933e-46
## age65-74     3.3505241  0.1847987  18.130666 1.825358e-73
## age75-84     3.7001281  0.1922190  19.249541 1.424180e-82
## smoke1       0.3545457  0.1073741   3.301966 9.600961e-04
```

As this model does not include an interaction term between age and smoking, we are assuming a constant ratio of death rates over the levels of age. As seen in **A**, there is a clear increase in death rate as age increases. Thus, this model does not seem to be entirely appropriate and we should consider an interaction term for smoking and age.

C Based on (a), explain why it is sensible to add a quantitative interaction of age and smoking. Specify this model, and show that the log of the ratio of coronary death rates changes linearly with age.

Here is the summary for our interaction term included model:

```
##                    Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)      -9.1480926  0.7071065 -12.937361 2.769814e-38
## age45-54          2.3575267  0.7637624   3.086728 2.023727e-03
## age55-64          3.8303228  0.7319248   5.233219 1.665833e-07
## age65-74          4.6228163  0.7319248   6.315971 2.684696e-10
## age75-84          5.2945191  0.7295600   7.257141 3.953588e-13
## smoke1            1.7470330  0.7288688   2.396910 1.653397e-02
## age45-54:smoke1 -0.9867826  0.7900623  -1.248993 2.116675e-01
## age55-64:smoke1 -1.3629685  0.7561867  -1.802423 7.147883e-02
## age65-74:smoke1 -1.4424497  0.7565317  -1.906661 5.656448e-02
## age75-84:smoke1 -1.8471513  0.7571735  -2.439535 1.470617e-02
```

For smokers, we have the following model:

$$log(\mu/t) = -9.1481 + (\beta_{age} + \beta_{age*smoke})Age + \beta_{smoke}$$

For nonsmokers, we have the following model:

$$log(\mu/t) = -9.1481 + (\beta_{age})Age$$

We can see that both log models includes a parameter related to age, and thus both models changes linearly with age.

D Fit the model in (b). Assign scores to the levels of age for a product interaction term between age and smoking, and fit the model in (c). Compare the fits by comparing the deviances. Interpret.

Here is a summary for the model including an interaction term of age scores with smoking status:

```
##                 Estimate Std. Error    z value      Pr(>|z|)
## (Intercept)   -8.5856465 0.30498334 -28.151198 2.316796e-174
## age45-54       1.7345703 0.21325066   8.133950  4.155223e-16
## age55-64       3.1483632 0.25334587  12.427135  1.861894e-35
## age65-74       4.1420286 0.31925255  12.974144  1.715012e-38
## age75-84       4.7309897 0.38396166  12.321516  6.937772e-35
## smoke1         1.4449920 0.37288584   3.875159  1.065550e-04
## smoke1:score  -0.3087396 0.09726458  -3.174224  1.502376e-03
```

The residual deviance for the model without an interaction term is 12.1339097 and the residual deviance for the model with the score interactive term is 1.5464376. We can use this ΔG^2 to conduct a test for the two models:

Hypothesis: H_0 : Model with no interaction VS H_a : Model with interaction

Test statistic: $\Delta g^2 = 10.5874721$

Distribution under null: χ^2 with df = 1

P-value: 0.0011386

Thus, we proceed to reject the numm hypothesis and conclude that the interaction term is beneficial to the overall model.

6.1

A model fit predicting preference for President (Democrat, Republican, Independent) using x = annual income (in 10,000 dollars) is $log(\pi^D/\pi^I)$ = 3.3 - 0.2x and $log(\pi^R/\pi^I)$ = 1.0 + 0.3x.

A

State the prediction equation for $log(\pi^R/\pi^D)$. Interpret its slope.

Given the previous equations, we have:

$$log(\pi^R/\pi^D) = log\Big(\frac{\pi^R/\pi^I}{\pi^D/\pi^I}\Big) = log(\pi^R/\pi^I) - log(\pi^D/\pi^I) = 1.0 + .3x - (3.3 - .2x) = -2.3 + .5x$$

As the slope here is 0.5, we can see that as income increases, voters are more likely to vote for a Republican candidate. Specifically the odds change by $e^{0.5} = 1.6487213$ for each unit increase in income.

B

Find the range of x for which $\pi^R > \pi^D$

Here, the probability inequality can be expressed as the inequality between the fitted models for predicting preference:

$$1.0 + .3x > 3.3 - .2x \rightarrow .5x > 2.3 \rightarrow x > 4.6$$

As x was provided in 10,000 dollars, this range is defined as an income greater than 46,000 dollars.

C

State the prediction equation for $\pi\hat{}I$:

For any value j in the model, we have the following prediction equation:

$$\frac{e^{\alpha_j+\beta_j x}}{\sum_{i=1}^{j} e^{\alpha_j+\beta_j x}}$$

Given the three outcomes and the baseline of Independent (all coeffecients are set to zero), we have

$$\hat{\pi}_I = \frac{e^{\alpha_I+\beta_I x}}{e^{\alpha_I+\beta_I x} + e^{\alpha_R+\beta_R x} + e^{\alpha_D+\beta_D x}} = \frac{e^{0+0x}}{e^{0+0x} + e^{1+.3x} + e^{3.3-.2x}} = \frac{1}{1 + e^{1+.3x} + e^{3.3-.2x}}$$

6.2

Refer to the alligator food choice example in Section 6.1.2.

A

Using the model fit,estimate an odds ratio that describes the effect of length on primary food choice being either invertebrate or other.

From the results, we can define the equation for invertebrates with a baseline of others as

$$log(\hat{\pi}_I/\hat{\pi}_O) = 5.6974 - 2.4654x$$

As the slope for this equation is -2.465, we are able to say that the odds of primary food choice being invertebrate as opposed to other are $e^{-2.465} = 0.0850088$ for each unit increase in length.

B

Estimate the probability that food choice is invertebrate, for an alligator of length 3.9 meters.

We can define the probability model for invertebrates as follows (similar to the model found in 6.1)

$$\hat{\pi}_I = \frac{e^{5.6974-2.4654x}}{e^{5.6974+-2.4654x} + e^{1.6177+-.1101x} + 1}$$

Using a length of 3.9, we can estimate the probability that the food choice will be invertebrates to be 0.005.

C

Estimate the length at which the outcomes invertebrate and other are equally likely.

We have defined the probability model for invertebrates as

$$\hat{\pi}_I = \frac{e^{5.6974-2.4654x}}{e^{5.6974+-2.4654x} + e^{1.6177+-.1101x} + 1}$$

Likewise the probability model for other (the baseline) is defined as

$$\hat{\pi}_O = \frac{1}{e^{5.6974+-2.4654x} + e^{1.6177+-.1101x} + 1}$$

As we are seeking the length x at which the probabilities are equal, we are left with the following equality:

$$e^{5.6974-2.4654x} = 1 \rightarrow 5.6974 - 2.4654x = 0 \rightarrow 5.6974 = 2.4654x \rightarrow x = 2.31$$

Thus, an allligator of length 2.31 meters will have equal probabilities of choosing invertabrates and others as their food choice.

6.3

Here is a summary of our model with a baseline of Other food choice of length >2.3 in lake Oklawha:

```
##                     Estimate Std. Error     z value     Pr(>|z|)
## (Intercept):1   1.898618526  0.6428159  2.95359621 3.140947e-03
## (Intercept):2   1.286819246  0.6656108  1.93329092 5.320037e-02
## (Intercept):3   1.042957824  0.7157853  1.45708201 1.450937e-01
## (Intercept):4  -0.847666030  1.1690571 -0.72508524 4.683997e-01
## size<2.3:1     -0.331550263  0.4482467 -0.73966028 4.595062e-01
## size<2.3:2      1.126654357  0.5049112  2.23139132 2.565522e-02
## size<2.3:3     -0.682813102  0.6514090 -1.04820949 2.945421e-01
## size<2.3:4     -0.962209993  0.7127053 -1.35008115 1.769900e-01
## lakeHancock:1  -0.820543109  0.7295607 -1.12470855 2.607126e-01
## lakeHancock:2  -3.416121029  0.8742587 -3.90744867 9.327585e-05
## lakeHancock:3  -2.036638452  0.9763628 -2.08594426 3.698367e-02
## lakeHancock:4   0.527782243  1.3018565  0.40540740 6.851781e-01
## lakeTrafford:1 -1.510715630  0.7531385 -2.00589343 4.486762e-02
## lakeTrafford:2 -1.325950122  0.7467027 -1.77574042 7.577571e-02
## lakeTrafford:3 -1.034334283  0.8401848 -1.23107946 2.182931e-01
## lakeTrafford:4  0.230258923  1.3004404  0.17706226 8.594595e-01
## lakeGeorge:1    0.005653103  0.7765134  0.00728011 9.941914e-01
## lakeGeorge:2   -0.931566219  0.7967114 -1.16926427 2.422972e-01
## lakeGeorge:3   -2.453218859  1.2937643 -1.89618685 5.793534e-02
## lakeGeorge:4    0.658860897  1.3685454  0.48143152 6.302098e-01
```

A

Fit a model to describe effects of length and lake on primary food choice. Report the prediction equations.

From our coefficients, we have the following prediction equations

$$log(\pi_F/\pi_O) = 1.89 - .33(Small) - .821(Hancock) - 1.51(Trafford) + .006(George)$$

$$log(\pi_I/\pi_O) = 1.28 + 1.13(Small) - 3.42(Hancock) - 1.32(Trafford) - .932(George)$$

$$log(\pi_R/\pi_O) = 1.04 - .683(Small) - 2.03(Hancock) - 1.03(Trafford) - 2.45(George)$$

$$log(\pi_B/\pi_O) = -.848 - .962(Small) + .528(Hancock) + .23(Trafford) + .659(George)$$

B

Using the fit of your model, estimate the probability that the primary food choice is fish for each length in Lake Oklawaha. Interpret the effect of length.

Considering the food choice of Fish and Lake of Oklawaha, we have the following probability model comparing fish to all other food choices:

$$\frac{e^{1.89-.33(Small)}}{1+e^{1.89-.33(Small)}+e^{1.28+1.13(Small)}+e^{1.04-.683(Small)}+e^{-.848-.962(Small)}}$$

Thus, the probability of food choice being fish from lake Oklawaha is 0.2581861 for small fish and 0.4584385 for large fish.

Appendix

1 Vehicle Safety

A Fit a baseline category logit model for this data using no/little importance as the baseline for the response. Interpret the parameters in the logit equation for important vs no/little importance.

```r
car = read.csv("car.csv",header=T)
car2 = car %>% spread(response,freq)
```

Here is a summary of our logit model, using the given baseline. In this instance, coeffecients with index **1** and **2** are related to Important and Very Important, respectively.

```r
model_1 = vglm(cbind(Important,Very.important, No.little.importance)~age+sex,data=car2, family=multinomi
summary = summary(model_1)
coef(summary)
```

Focusing on important vs little/no importance, we have the following log odds equation:

$$log(\pi_I/\pi_{NLI}) = .6087 - 1.5877(Age_{18-23}) - .459(Age_{24-40}) + .388(Sex_F)$$

i Estimate the probability of no/little importance for women aged 18-23.

```r
#predict(model_1, type = "response")
```

The probability of no/little importance for women ages 18-23 is defined as

$$\frac{1}{1 + exp(.608 - 1.588 + .388) + exp(1.06 - 2.916 + .813)}$$

```r
pi.i = exp(0)/(1 + exp(coef(model_1)[1] + coef(model_1)[3] + coef(model_1)[7])
                 + exp(coef(model_1)[2] + coef(model_1)[4] + coef(model_1)[8]))
```

ii Estimate the probability of no/little importance for men age greater than 40.

The probability of no/little importance for men age greater than 40 is defined as

$$\frac{1}{1 + exp(.608) + exp(1.06)}$$

```r
pi.ii = exp(0)/(1 + exp(coef(model_1)[1])
                 + exp(coef(model_1)[2]))
```

B Fit a proportional odds models. Interpret the parameters in one of your equations.

```r
response1 = relevel(car$response,ref="No.little.importance")
sex1 = relevel(car$sex,ref="women")
age1 = relevel(car$age,ref="18-23")
POM_1 =  polr(response1 ~age1+sex1, data=car,weight = freq)
```

Here is the summary for our proportional odds model:

```
summary = summary(POM_1)
coef(summary)
```

i Find the probability of each category for females aged 18-23.

```
pi.less.N = exp(POM_1$zeta[1])/(1 + exp(POM_1$zeta[1]) )

pi.less.I = exp(POM_1$zeta[2])/(1 + exp(POM_1$zeta[2]) )
```

```
probabilities = predict(POM_1,newdata = data.frame(age1="18-23",sex1="women"), type = "probs")
probabilities
```

2 Car Accidents

A Considering a male wearing a seatbelt with an accident in an urban location as the baseline. Fit a proportional odds model. State the model and the distribution assumptions.

```
accidents = data.frame(expand.grid(belt=c("No","Yes"),
                                   location=c("Urban","Rural"),
                                    gender=c("Female","Male"),
                                   response=c("one","two","three","four","five")),
        count=c(7287,11587,3246,6134,10381,10969,6123, 6693,
        175,126,73,94,136,83,141,74,
        720,577,710,564,566,259,710,353,
        91,48,159,82,96,37,188,74,
        10,8,31,17,14,1,45,12))
```

```
accidents$gender = relevel(accidents$gender,ref = "Male")
accidents$belt = relevel(accidents$belt,ref = "Yes")
accidents$location = relevel(accidents$location,ref = "Urban")

POM_2 =  polr(response ~gender+belt+location, data=accidents,weight=count)
```

B Given the baseline, find the estimated probabilities for each of the response categories.

```
pi.less.1 = exp(POM_2$zeta[1])/(1 + exp(POM_2$zeta[1]) )
pi.less.2 = exp(POM_2$zeta[2])/(1 + exp(POM_2$zeta[2]) )
pi.less.3 = exp(POM_2$zeta[3])/(1 + exp(POM_2$zeta[3]) )
pi.less.4 = exp(POM_2$zeta[4])/(1 + exp(POM_2$zeta[4]) )
pi.less.5 = 1
```

Using the **predict** function, here are the estimated probabilties for each response given the input of Males wearing seatbelts in an Urban area:

```
probabilities = predict(POM_2,newdata = data.frame(gender="Male",belt="Yes",location="Urban"), type = "
probabilities
```

We can see that the predict function and the manual computations are the same.

C Find the cumulative odds ratio of gender, given location and seatbelt use. Interpret this odds ratio.

The odds ratio for gender in this scenario is defined as

$$\frac{odds(Y=i|female)}{odds(Y=i|male)}$$

```r
odds_coef = - POM_2$coefficients[1]
exp_odds_coef = exp(odds_coef)
```

D Find the cumulative odds ratio of seatbelt use, given that the accident occurred in a rural location. Does this differ from accidents in urban locations? If so, why?

```r
odds_belt_rural = -POM_2$coefficients[2]

exp_odds_belt_rural = exp(odds_belt_rural)
```

3 Estimate a Poisson regression model to predict the number of satellites as a function of the given explanatory variables. What should your final model be? Interpret the parameters in your final model. Give a summary of the process you used to determine this model.

```r
crab=read.table("crab.txt")
crab=crab[,-1]
colnames(crab)=c("C","S","W","Wt","Sa")
```

Here is the summary for the model explaining the number of satellites as a function of Width:

```r
model_3 = glm(Sa ~W, data = crab, family=poisson(link=log))
summary = summary(model_3)
coef(summary)
```

```r
model_3b = glm(Sa ~  W + C, data = crab, family=poisson(link=log))
summary = summary(model_3b)
coef(summary)
```

```r
#From table 3.3
width=c(22.125,23.75,24.75,25.75,26.75,27.75,28.75,31.4)
cases=c(14,14,28,39,22,24,18,14)
Sa=c(14,20,67,105,63,93,71,72)
Crab2=data.frame(width,cases,Sa)
Crab2
```

Here are the results from our grouped rate regression model:

```r
model_3c=glm(Sa~width + offset(log(cases)),family=poisson(link=log),data=Crab2)
summary(model_3c)
```

3.20

```
deaths = c(2,12,28,28,31,32,104,206,186,102)
personyears = c(18793,10673,5710,2585,1462,52407,43248,28612,12663, 5317)
smoke = factor(rep(0:1,c(5,5)))
age = factor(rep(c("35-44","45-54","55-64","65-74","75-84"),2))
three_20 = data.frame(deaths,personyears,smoke,age)
```

A For each age, compute the sample coronary death rates per 1000 person-years, for nonsmokers and smokers. To compare them, take their ratio and describe its dependence on age.

Here is our original table with a new column representing the sample death rate per 1000 person-years:

```
three_20$rate = 1000/three_20$personyears * three_20$deaths
three_20
```

We can observe that smokers have a higher rate of death for younger ages, but the ratio becomes approximately one for older ages.

B Specify a main-effects Poisson model for the log rates having four parameters for age and one for smoking. Explain why this model assumes a constant ratio of nonsmokers' to smokers' coronary death rates over levels of age. Based on (a), would you expect this model to be appropriate?

Here is the summary for our main-effects Poisson model:

```
model_3.20 = glm(deaths ~ age + smoke, data = three_20, family=poisson(link=log),offset = log(personyea
summary = summary(model_3.20)
coef(summary)
```

C Based on (a), explain why it is sensible to add a quantitative interaction of age and smoking. Specify this model, and show that the log of the ratio of coronary death rates changes linearly with age.

Here is the summary for our interaction term included model:

```
model_3.20_c = glm(deaths ~ age + smoke + age*smoke, data = three_20, family=poisson(link=log),offset =
summary = summary(model_3.20_c)
coef(summary)
```

D Fit the model in (b). Assign scores to the levels of age for a product interaction term between age and smoking, and fit the model in (c). Compare the fits by comparing the deviances. Interpret.

```
three_20$score = rep(c(1,2,3,4,5),2)
```

Here is a summary for the model including an interaction term of age scores with smoking status:

```
model_3.20_d = glm(deaths ~ age + smoke + smoke*score, data = three_20, family=poisson(link=log),offset
summary = summary(model_3.20_d)
coef(summary)
```

6.1

A model fit predicting preference for President (Democrat, Republican, Independent) using x = annual income (in 10,000 dollars) is $log(\pi^D/\pi^I)$ = 3.3 - 0.2x and $log(\pi^R/\pi^I)$ = 1.0 + 0.3x.

6.2

Refer to the alligator food choice example in Section 6.1.2.

B

Estimate the probability that food choice is invertebrate, for an alligator of length 3.9 meters.

We can define the probability model for invertebrates as follows (similar to the model found in 6.1)

```
food_prob = function(x){
  top = exp(5.6974 - 2.4654*x)
  bottom = exp(5.6974 - 2.4654*x) + exp(1.6177 - .1101*x) + 1
  prob = top/bottom
  return(round(prob,3))
}

three_nine = food_prob(3.9)
```

Using a length of 3.9, we can estimate the probability that the food choice will be invertebrates to be 0.005.

6.3

```
gator = data.frame(expand.grid(size=c("<2.3",">2.3"),
        lake=c("Hancock","Oklawaha","Trafford","George"),
        response=c("Fish","Invertebrate","Reptile","Bird","Other")),
        count=c(23,7,5,13,5,8,16,17,4,0,11,8,11,7,19,1,2,1,1,6,2,6,1,0,2,3,0,1,1,3,2,1,8,5,3,0,5,5,3,3)

gator$lake = relevel(gator$lake,ref = "Oklawaha")
gator$size = relevel(gator$size,ref = ">2.3")
gator$response = relevel(gator$response,ref = "Other")

gator_6.3 = gator %>% spread(response, count)
```

Here is a summary of our model with a baseline of Other food choice of length >2.3 in lake Oklawha:

```
model_6.3 = vglm(cbind(Fish,Invertebrate,Reptile,Bird,Other)~size +lake,data=gator_6.3, family=multinom
summary = summary(model_6.3)
coef(summary)
```

B

```
pi.small = exp(coef(model_6.3)[1]+coef(model_6.3)[5])/(1+ exp(coef(model_6.3)[1]+coef(model_6.3)[5])
                                                       + exp(coef(model_6.3)[2]+coef(model_6.3)[6])
                                                       + exp(coef(model_6.3)[3]+coef(model_6.3)[7])
                                                       + exp(coef(model_6.3)[4]+coef(model_6.3)[8]))
```

```
pi.large = exp(coef(model_6.3)[1])/(1+ exp(coef(model_6.3)[1])
                                     + exp(coef(model_6.3)[2])
                                     + exp(coef(model_6.3)[3])
                                     + exp(coef(model_6.3)[4]))
```