Project #4

STAT 875

Spring 2014

Complete the following problems below. Within each part, include your R program output with code inside of it and any additional information needed to explain your answer. Your R code and output should be formatted in the same manner as in the lecture notes.

1. (34 total points) Pneumoconiosis (black lung disease) “is a lung disease that results from breathing in dust from coal, graphite, or man-made carbon over a long period of time” (<http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001187>). The data frame pneumo in the VGAM package contains counts of coal miners aggregated into eight groups based on their time working in the mine (in years) and their severity of pneumoconiosis (three levels: normal, mild, and severe). Below is the data:

> library(VGAM)

> pneumo

 exposure.time normal mild severe

1 5.8 98 0 0

2 15.0 51 2 1

3 21.5 34 6 3

4 27.5 35 5 8

5 33.5 32 10 9

6 39.5 23 7 8

7 46.0 12 6 10

8 51.5 4 2 5

Complete the following:

* 1. (3 points) The data set is in a different form than we have seen so far in Chapter 3, but a multinomial regression model can still be estimated using:

mod.fit.nom1<-multinom(formula = cbind(normal, mild, severe) ~ exposure.time, data = pneumo)

State the estimated model. Try to perform a LRT using Anova() to determine the importance of exposure.time. What problems occur?

* 1. (2 points) Transform the data set to the usual form with the following code:

pneumo.normal<-data.frame(time = pneumo$exposure.time, count = pneumo$normal,

 resp = "normal")

pneumo.mild<-data.frame(time = pneumo$exposure.time, count = pneumo$mild, resp =

 "mild")

pneumo.severe<-data.frame(time = pneumo$exposure.time, count = pneumo$severe, resp

 = "severe")

set1<-rbind(pneumo.normal, pneumo.mild, pneumo.severe)

Describe what each line of code does. Print the first six observations. Use this form of the data set for the rest of the project.

* 1. (2 points) Would a multinomial or proportional odds regression model typically be better to use with this type of data? Explain. This explanation should be based on the structure of the data rather than the subsequent statistical analysis that you will be performing.
	2. For a multinomial regression model:
		1. (2 points) Estimate a multinomial regression model using time as the explanatory variable. Verify the model is the same as found in part a).
		2. (2 points) Assess the importance of exposure time using a LRT. Make sure to properly state the hypotheses and use α = 0.05.
		3. (2 points) Estimate the probabilities for the pneumoconiosis categories for each age group and display these probabilities in an appropriate data frame.
		4. (3 points) Plot the estimated model. Describe the relationship between pneumoconiosis and exposure time.
		5. (4 points) Estimate the appropriate odds ratios to fully understand the relationships between exposure time and pneumoconiosis severity. Calculate 95% Wald intervals as well. Fully interpret the odds ratios in the context of the data. Use a 5 year increase in exposure time for your calculations.
	3. (10 points) For a proportional odds model, complete the same items as in part d). For the plot part, include your model on the same plot as in part d) (i.e., include the multinomial regression model on the plot). For the odds ratio part, use profile LR intervals.
	4. (2 points) Choose the appropriate question to answer below.
		1. Based on your statistical analysis, what is gained by using the multinomial model rather than the proportional odds model?
		2. Based on your statistical analysis, what is gained by using the proportional odds model rather than the multinomial model?
	5. (2 points) What can you conclude about the relationship between pneumoconiosis and exposure time?
1. (5 points) Suppose a multinomial regression model has one continuous explanatory variable x that is represented in the model by both a linear and quadratic term. For a c-unit increase in x, derive the odds ratio that compares a category j (j ≠ 1) response to a category 1 response for j = 2, …, J (J > 2). Show the form of the estimated variance (expressed in individual terms that can be obtained directly from the covariance matrix) that would be used in a Wald confidence interval.