MAST90139: Statistical Modelling for Data Science Assignment 3

Due time: 5 pm Friday May 27, 2022.

1. Through radiological examination 371 coal miners were classified into 3 categories of pneumonoconiosis: `1normal`, `2mild` and `3severe`. These coal miners were also classified into 8 groups according to the number of years each had spent working at the coal face. Data summarizing these cross classifications are given in the data frame `pneumo` as shown below:

   [10]

```
> pneumo

   Freq  status year
1    98 1normal  5.8
2    51 1normal 15.0
3    34 1normal 21.5
4    35 1normal 27.5
5    32 1normal 33.5
6    23 1normal 39.5
7    12 1normal 46.0
8     4 1normal 51.5
9     0   2mild  5.8
10    2   2mild 15.0
11    6   2mild 21.5
12    5   2mild 27.5
13   10   2mild 33.5
14    7   2mild 39.5
15    6   2mild 46.0
16    2   2mild 51.5
17    0 3severe  5.8
18    1 3severe 15.0
19    3 3severe 21.5
20    8 3severe 27.5
21    9 3severe 33.5
22    8 3severe 39.5
23   10 3severe 46.0
24    5 3severe 51.5
```

Treating pneumonoconiosis `status` as a nominal categorical response variable, a multi-categorical logit model has been fitted resulting in the following `R` output:

```
> pneumo$status=relevel(pneumo$status, ref="1normal")
> nominal.mod <- multinom(status~year, data=pneumo, weights=Freq, Hess=T)
> summary(nominal.mod)

Coefficients:
        (Intercept)       year
2mild     -4.2917        0.0836
3severe   -5.0598        0.1093

Std. Errors:
        (Intercept)       year
2mild      0.5214        0.0153
3severe    0.5964        0.0165
```

(a) Write down the model fitted in the above `R` output. You need to define the response variable and covariate for the model. Also, you need to specify the probability distribution of the response variable and estimates of all parameters in the model.

(b) Provide an interpretation for the coefficient estimate 0.1093. Then calculate an approximate 95% confidence interval for the odds ratio of severe status versus normal status for every 10 more years spent working at the coal face.

(c) Estimate the pneumonoconiosis status probabilities for a miner who has spent 25 years working at the coal face.

2. Refer to the `pneumo` data in Q1. Treating the pneumonoconiosis `status` as an ordinal categorical response variable, a cumulative model is fitted producing the following `R` output (Note `Coefficients Value` needs to change sign for being used in the model): `[10]`

```
> pneumo$status=as.ordered(as.character(pneumo$status))
> ordinal.mod=polr(status~year, data=pneumo,weights=Freq, Hess=T, method="logistic")
> summary(ordinal.mod)

Coefficients:
      Value Std. Error t value
year 0.0959    0.01194   8.034

Intercepts:
              Value   Std. Error t value
1normal|2mild  3.9558  0.4097     9.6558
2mild|3severe  4.8690  0.4411    11.0383
```

(a) Write down the model fitted in the above `R` output. You need to define the response variable and covariate for the model. Also, you need to specify the probability distribution of the response variable and estimates of all parameters in the model.

(b) Provide an interpretation for the coefficient estimate 0.0959. Then calculate an approximate 95% confidence interval for the odds ratio of non-normal status versus normal status for every 10 more years spent working at the coal face.

(c) Estimate the pneumonoconiosis status probabilities for a miner who has spent 25 years working at the coal face.

3. The `ohio` data concern 537 children from Steubenville, Ohio and were taken as part of a study on the effects of air pollution. Children were in the study for four years from age seven to ten. The response is whether they wheezed or not. The variables are [10]

   resp: an indicator of wheeze status (1=yes, 0=no)
   id: an identifier for the child, taking values from 0 to 536
   age: 7 yrs = -2; 8 yrs = -1; 9 yrs =0; 10 yrs =1
   smoke: mother's smoking status at start of the study (1=smoker, 0=nonsmoker)

   Some analysis has been done to the data in `R`, producing the following output.

```
> head(ohio)

  resp id age smoke
1    0  0  -2     0
2    0  0  -1     0
3    0  0   0     0
4    0  0   1     0
5    0  1  -2     0
6    0  1  -1     0

> tail(ohio)

     resp  id age smoke
2143    1 535   0     1
2144    1 535   1     1
2145    1 536  -2     1
2146    1 536  -1     1
2147    1 536   0     1
2148    1 536   1     1

> str(ohio)

'data.frame':   2148 obs. of  4 variables:
 $ resp : int  0 0 0 0 0 0 0 0 0 0 ...
 $ id   : int  0 0 0 0 1 1 1 1 2 2 ...
 $ age  : int  -2 -1 0 1 -2 -1 0 1 -2 -1 ...
 $ smoke: int  0 0 0 0 0 0 0 0 0 0 ...
```

```
> library(geepack}
> fit.exch <- geeglm(resp~age+smoke, family=binomial(link="logit"),
data=ohio, id=id, corstr = "exchangeable", std.err="san.se"); summary(fit.exch)

 Coefficients:
            Estimate  Std.err    Wald Pr(>|W|)
(Intercept) -1.880    0.114   272.597  < 2e-16 ***
age         -0.113    0.044     6.684  0.00973 **
smoke        0.265    0.178     2.224  0.13588

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)   0.9985  0.1116

Correlation: Structure = exchangeable  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha   0.3543 0.06244
Number of clusters:   537   Maximum cluster size: 4
```

Use the above output to answer the following questions.

(a) Let y_{it} be the response value `resp` of child i at age t. Write down the model involved in the analysis, including the mean, variance and correlation coefficient of y_{it}'s. Give the estimates of the parameters appearing in the model.

(b) Write down the model's design matrix for data where `id`=536.

(c) Estimate the odds ratio of wheezing for a child for every one year older in age. Also calculate an approximate standard error of your odds ratio estimate.

(d) Estimate the odds ratio of wheezing for a 10-year old child with a smoking mother versus a 9-year old child with nonsmoking mother.

`Total marks = 30`