Alex Jones

Molstad

STA4504

4/21/2022

STA4504 Exam 2

1.

a) Creating the Baseline-Category Logit model in R, we produce the following output:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept):1   1.0986     0.4364   2.517  0.01183 *
(Intercept):2   0.1335     0.5175   0.258  0.79640
Events10-14:1  -1.5731     0.4984  -3.156  0.00160 **
Events10-14:2  -0.5062     0.5677  -0.892  0.37260
Events15+:1    -1.6805     0.5220  -3.219  0.00129 **
Events15+:2    -0.5688     0.5854  -0.972  0.33121
Events5-9:1    -0.2513     0.5345  -0.470  0.63824
Events5-9:2     0.1542     0.6201   0.249  0.80368
---
```

Note that the reference group for this model is the "Weakened" response, and thus does not receive a linear model. Looking at the output above, we produce the two following equations for our model:

$$\log\left(\frac{\pi_1}{\pi_3}\right) = 1.0986 - 0.2513I_{5-9} - 1.5731I_{10-14} - 1.6805I_{15+}$$

$$\log\left(\frac{\pi_2}{\pi_3}\right) = 0.1334 + 0.1542I_{5-9} - 0.5062I_{10-14} - 0.5688I_{15+}$$

Where $\pi_1$, $\pi_2$, and $\pi_3$ denote the probabilities of symptoms being improved, unchanged, or worsened respectively.

b) To do a significance test, we will perform a likelihood ratio test comparing the model with Events to one without. Our null hypothesis is $H_0: \beta_{11} = \beta_{12} = \beta_{13} = \beta_{21} = \beta_{22} = \beta_{23} = 0$, with our alternative hypothesis being that $\beta_{ij} \neq 0$ for some $i = 1, 2$ and $j = 1, 2, 3$. Creating the model without events and performing a likelihood ratio test on them, our test-statistic comes out to be 24.677, which on a null chi-squared distribution with 6 degrees of freedom, gives us a p-value of 0.0003918769. As such, we can reject the null hypothesis, suggesting that we have sufficient evidence to include events in our model, and that there appears to be an association between events and symptoms.

c) Because the interval asked is the opposite of the first equation, we will create a confidence interval for that model, and then take the negative. Doing this, the point estimate is $\alpha_{11} + \beta_{13} = 1.0986 - 1.6805 = -0.5819$. Constructing our standard error while taking covariance into account, SE comes out to be 0.286432. Using the formula for the 95% Wald Confidence Interval:

$$\bar{x} \pm 1.96(SE)$$

Our initial interval comes out to be (-1.1433, -0.0204), and our final interval for $\log\left(\frac{P(Symptoms = Worsened \mid Events = 15+)}{P(Symptoms = Improved \mid Events = 15+)}\right)$ comes out to be (0.0204, 1.1433). We are 95% confident that the true log-odds-ratio between symptoms being worsened vs. symptoms being improved given 15+ flare-up events is captured in the interval (0.0204, 1.1433) In the context of the study, this means that when events are greater than 15, worsened symptoms are more likely to be observed than improved.

d) Yes, the model is still valid.

e) Adding ordinality into the model would be beneficial because it's another piece of information that we know in the relationship between events and symptoms. It will better help us investigate the trend between flare-up events and symptom levels.

f) Using R to produce the cumulative logit (which uses ordinality), we produce the following output:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept):1   0.2950     0.2332   1.265    0.206
(Intercept):2   1.5507     0.2461   6.302 2.93e-10 ***
Events10-14    -1.2850     0.2596  -4.950 7.41e-07 ***
Events15+      -1.3678     0.2716  -5.037 4.74e-07 ***
Events5-9      -0.2819     0.2745  -1.027    0.304
---
```

Again, we have equations for the model, written as such:

$$logit\big(P(Y \leq 1)\big) = 0.2950 - 0.2819I_{5-9} - 1.2850I_{10-14} - 13.678.\, I_{15+}$$
$$logit\big(P(Y \leq 2)\big) = 1.5507 - 0.2819I_{5-9} - 1.2850I_{10-14} - 13.678.\, I_{15+}$$

g) Like 1b), we will be performing a likelihood ratio test comparing this model to one without events as a predictor. For this, our null hypothesis will be $H_0: \beta_1 = \beta_2 = \beta_3 = 0$, with alternative hypothesis $\beta_i \neq 0$ for some $i = 1, 2, 3$. Constructing the model without events and comparing deviances, we produce a test-statistic of 24.43, which when following the null chi-squared distribution with 3 degrees of freedom, gives us a p-value of 2.032e-05. As such, we can reject the null hypothesis, suggesting that we have sufficient evidence to include events in our model, and that there appears to be an association between events and symptoms.

h) We will construct the confidence similar to how we did it in 1c), however since the interval is for the odds (and not log-odds), we will negate the value and then exponentiate. With a point estimate of 0.1829 (derived from using the second equation) and computed standard error of 0.1058978, we compute the initial 95% Wald Interval to be (-0.02465965, 0.3904597), and after doing the referenced transformations, the proper interval becomes (0.6767, 1.0250). We can say with 95% confidence that the true odds-ratio between symptoms being worsened vs. improved or unchanged is within the interval (0.6767, 1.0250).

i) To find the estimate using the given model, a few things important to note are that the equation with Intercept 2 can be converted to $P(Y \leq 2|\text{Events} = 15+) = \frac{e^{\alpha_2+\beta_{15+}}}{1+e^{\alpha_2+\beta_{15+}}}$,

$P(Y \leq 1|\text{Events} = 15+)$ can be found similarly using intercept 1 instead, and that $P(Y = 2) = P(Y \leq 2) - P(Y \leq 1)$. Using these properties:

$$P(Y = 2|\text{Events} = 15+) = \frac{e^{\alpha_2+\beta_{15+}}}{1+e^{\alpha_2+\beta_{15+}}} - \frac{e^{\alpha_1+\beta_{15+}}}{1+e^{\alpha_1+\beta_{15+}}}$$
$$= \frac{e^{0.57963-0.07639}}{1+e^{0.57963-0.07639}} - \frac{e^{-0.77309-0.07639}}{1+e^{-0.77309-0.07639}}$$
$$= 0.62322 - 0.29951$$
$$= 0.32368$$

j) Creating the model that adds ordinality to the events predictor, we produce the following output:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1  0.84680    0.23816   3.556 0.000377 ***
(Intercept):2  2.08826    0.25788   8.098 5.59e-16 ***
EventsNum     -0.52269    0.08034  -6.506 7.73e-11 ***
---
```

Testing for independence between events and symptoms, we will again perform a likelihood ratio test similar to the one in 1g). We will have null hypothesis $H_0$: $\beta = 0$ vs $\beta \neq 0$, where $\beta$ denotes the effect flare-up events have on symptom response. Reusing an earlier constructed model that doesn't contain flare-up events as a predictor and comparing deviances, we produce a test statistic of 20.711, which when following the null chi-squared distribution with 1 degree of freedom, gives us a p-value of 5.342e-06. As such, we can reject the null hypothesis, giving us sufficient evidence that there is a link between flare-up events and symptom response, and that they are not independent of each other.

k) The two models are not nested, so a likelihood ratio test cannot be done. AIC or BIC should be used to compare the models instead.

l) Comparing AIC values, the model I will select is the last one, which takes ordinality into account for both flare-up events and symptom response. This is because this model had the lowest AIC of the three, at 608.2998. Looking at the model, as more flare-ups occur, the efficacy of the treatment diminishes, to the point where if a patient experiences fifteen or more flare-ups over the three-month period, their symptoms are more likely to worsen than to stay the same or improve.

2.

a) Creating a GEE model with independent correlation, we produce the following output:

```
Coefficients:
             Estimate Naive S.E.  Naive z Robust S.E.  Robust z
(Intercept) -1.0269267 0.23024541 -4.46014  0.20730465 -4.953708
severity     0.2939677 0.04339065  6.77491  0.04011126  7.328809
```

We still have valid results regardless of which structure we choose; the only differences will be the robust standard errors.

b) Because the correlation matrix is between observations at different points in time, it seems best that an auto-regressive correlation structure is used. Constructing such a model in R, we produce the following output:

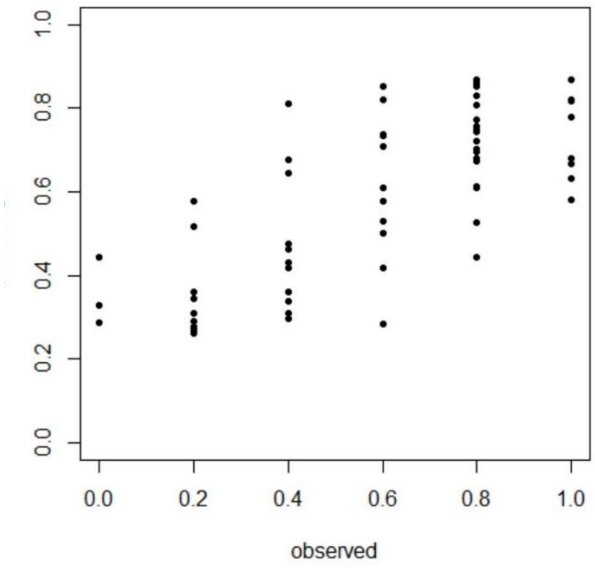
```
Coefficients:
             Estimate Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept) -1.0272332 0.23506279 -4.370038  0.20689509 -4.964996
severity     0.2941252 0.04430371  6.638838  0.04007146  7.340016
```

Considering how close the estimates of the auto-regressive model are to the independent model, it seems that checking for association between time observed is unnecessary in the context of this study.

c) Performing a significance test on the association between symptom severity and presence of side effects, we will use the null hypothesis $H_0: \beta = 0$ vs $H_A: \beta \neq 0$. Using the robust standard errors and output from the GEE model, we find our test-statistic to be 7.340016, which corresponds to a p-value of 2.136069e-13. As such, we reject the null hypothesis, finding that we have sufficient evidence to say that the severity of subject's RA is associated with the probability that symptoms appear.

d) For the GLMM to match the GEE model, there must be an assumption that there is no subject-specific effect in play when comparing RA severity and probability of side-effects appearing.

e) Creating the GLMM that includes the subject-specific events, and then plotting the fitted values from the model opposite the observed values, we produce the following plot:




One thing to note is that range of proportions that are observed go from [0,1] while the range of predicted probabilities only go from [0.2626840, 0.8673733]. This shows that the GLMM is taking extreme values and pushing them towards a less varied probability. Another thing to note is despite many subjects having the same

proportion of weeks with flare-ups, the model predicts different probabilities for different subjects, indicating that there is a subject-specific effect that is occurring.

f) Performing the likelihood ratio test using the GLMM, we will be comparing the current model to one without severity as a predictor. As such, our null hypothesis is $H_0: \beta = 0$ vs. $H_A: \beta \neq 0$. Creating the model that neglects severity, we compare the two deviances and our test statistic comes out to be 40.744, which on the null chi-squared distribution with 1 degree of freedom, gives us a p-value of 1.735e-10. Due to such a low p-value, we choose to reject the null hypothesis, concluding that we have significant evidence to conclude that there is a link between severity of RA observed and whether side-effects will appear.

g) No, they cannot be interpreted the same. Since $\beta^*$ is for the GLMM, it can be interpreted as the change in log-odds of a side-effect appearing once the subject specific effect has been accounted for. $\beta$ does not account for subject specific effects, and thus cannot be interpreted the same.

h) In events where we do not care about subject-specific effects, a GEE will generally be preferred over a GLMM. A GEE will also check for association between predictor levels (in this example, weeks of check-up) which a GLMM will not specifically account for.

3.

a) Looking at the given graph, there are four major log-linear models that can be constructed:

   i) (AC, AD, BD, CD, CE, CF, EF)

   ii) (ACD, BD, CE, CF, EF)

   iii) (AC, AD, BD, CD, CEF)

   iv) (ACD, BD, CEF)

b) A and F are conditionally independent give C, since C is the node that the two are connected to that creates a path between them.

c) B and F are not independent. If either of them were independent from the other, a path could not be drawn leading one to the other, but since the path {B, D, C, F} exists, this tells us that they are not completely independent of each other. However, they are conditionally independent given {D,C}.

d) After collapsing Model 2 into Model 1, since the connection between E and F are undisturbed by neglecting {AC, DC, DB}, the estimate of the marginal association will remain the same.

e) 15 total parameters are needed to specify the log-linear model corresponding to this graph:
   6 are needed for each node A-F
   7 are needed for each edge (AC, AD, BD, CD, CE, CF, EF)
   2 are needed for the 3-cycles (ACD, CEF)

f) Creating the homogenous association model in R, we collect the following output:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.5439     0.3484   4.432 9.35e-06 ***
A2            0.5705     0.3886   1.468  0.14208
A3            0.5396     0.3857   1.399  0.16185
C2            0.5744     0.3295   1.743  0.08129 .
D2            1.0229     0.3906   2.619  0.00882 **
D3            1.6866     0.3657   4.612 3.99e-06 ***
A2:C2         0.2405     0.2903   0.829  0.40734
A3:C2         0.4762     0.2899   1.642  0.10049
A2:D2        -0.7200     0.4206  -1.712  0.08695 .
A3:D2        -0.7305     0.4117  -1.774  0.07601 .
A2:D3        -0.7806     0.3956  -1.973  0.04846 *
A3:D3        -0.8683     0.3896  -2.228  0.02585 *
C2:D2        -0.1980     0.3371  -0.587  0.55690
C2:D3        -0.5582     0.3125  -1.786  0.07412 .
---
```

This means that our log-linear model is:

$$\log(\hat{\mu}_{ijk}) = 1.5439 + 0.5705I_{2++} + 0.5396I_{3++} + 0.5744I_{+2+} + 1.0299I_{++2} + 1.6866I_{++3} + 0.2405I_{22+} + 0.4762I_{32+} - 0.7200I_{2+2} - 0.7305I_{2+3} - 0.7806I_{3+2} - 0.8683I_{3+3} - 0.1980I_{+22} - 0.5582I_{+23}$$

(note: the order of subscripts in $I$ are A,C,D)

g) One assumption that we made is that we have independent samples being taken, and that the interaction terms describing the homogenous association is sufficient.

h) Testing the latter, we will perform a likelihood ratio test comparing the homogenous association model with the saturated model. As such, our null hypothesis is $H_0$: the homogenous association model is sufficient vs. $H_A$: the model is not sufficient to estimate counts. Constructing the saturated model and performing a likelihood ratio test comparing the deviances, our test statistic comes out to be 1.4051, which under the null chi-squared distribution with 4 degrees of freedom, gives us a p-value of 0.8433. Given such a high p-value, we fail to reject the null hypothesis, and conclude that the homogenous association model is sufficient in estimating counts.

i) Testing whether A, C, and D are independent, we will perform a likelihood ratio test comparing the independent model with the saturated model. As such, our null hypothesis is $H_0$: the independence model is sufficient to estimate counts vs $H_A$: the model is not sufficient to estimate counts. Constructing the independence model and comparing the deviances, our test statistic comes out to be 15.841, which under the null chi-squared distribution with 12 degrees of freedom, gives us a p-value of 0.1986. As such, we fail to reject the null hypothesis, and conclude that the independence model is sufficient to estimate counts.

j) Finding the point estimate for the odds-interval ratio:

$$\frac{\hat{\mu}_{3c1}\hat{\mu}_{2c2}}{\hat{\mu}_{3c2}\hat{\mu}_{2c1}} = \exp(\lambda_{31}^{AD} + \lambda_{22}^{AD} - \lambda_{32}^{AD} - \lambda_{12}^{AD})$$

$$= \exp(\lambda_{22}^{AD} - \lambda_{32}^{AD})$$
$$= \exp(-0.7200 - (-0.7806))$$

Our point estimate comes out to $e^{0.0606}$. Using the vcov() function to construct the standard error of our estimate:

$$Var(\lambda_{22}^{AD} - \lambda_{32}^{AD}) = Var(\lambda_{22}^{AD}) + Var(\lambda_{32}^{AD}) - 2Cov(\lambda_{22}^{AD}, \lambda_{32}^{AD})$$
$$= 0.176928465 + 0.169513746 - 2(0.108850699)$$
$$= 0.12874$$
$$SE = 0.3588$$

Our interval can then be constructed as $e^{0.0606 \pm 1.96(0.3588)} = (0.74215, 1.52106)$. This means we are 95% confident that the odds ratio of observing $D = 1$ over $D = 2$, when conditioned on $A = 3$, is between 0.742 and 1.521 times the odds ratio of the same thing when conditioned on $A = 2$. The quantity can be rewritten as:

$$\frac{\left(\frac{P(D = 1|C = c, A = 3)}{P(D = 2|C = c, A = 3)}\right)}{\left(\frac{P(D = 1|C = c, A = 2}{P(D = 2, C = c, A = 2)}\right)}$$

Which affirms the interpretation of this interval.

k) Yes, it would be different. The cross-interaction terms for each level of A,C,D would be different for the saturated model so it would produce different results.